Responsible AI Guide

Coalition for Health AI (CHAI)

Version Table		
v0.1	*Draft 1; feedback incorporated from Editorial/Workgroup Leads, released to Independent Review Group*	*5/8/2024*
v0.2	*Draft 2; feedback incorporated from Independent Review Group and HHS/ONC, released to Board of Directors*	*6/14/2024*
v0.3	*Draft 3; feedback incorporated from Board of Directors, released for public comment*	*6/26/2024*
v0.4	*Draft 4; feedback incorporated from public comment*	*TBD*
v1.0	*Final Public Release*	*TBD*

Editorial Note

This Guide aims to be comprehensive in scope as of its date of publication, incorporating best practices for anyone responsible for or impacted by the development and deployment of AI solutions in healthcare. Future editions of this Guide or companion documents will meet the following aims:

1. Identify stakeholder-based actions and pathways across the AI lifecycle. As consensus develops across CHAI stakeholder groups (see Sections 3.2 and 3.3), future versions of the Guide will more crisply identify who is responsible for what actions at each stage of an AI solution's development and deployment.

2. Differentiate processes AI models developed, fine-tuned, or extended internally versus those adopted from external partners by implementing organizations. Future versions of the Guide will consider AI solutions that have already been developed and are entering a new healthcare environment at the "Assess" stage of the AI lifecycle (see Stage 4 in Section 4 below).

3. Clearly illustrate which considerations fall within which existing regulations. This version of the Guide is broad in the sense that it covers best practices already under the regulatory guidance of the FDA, ONC, and other entities. While it cites existing rules, this edition does not specifically outline the overlaps between best practices and existing regulations. Future versions of this Guide will more clearly define these overlaps.

Contributors

Principal Writing Team: Nicoleta Economou, Matthew Elmore, Alison Callahan, Jonathan McCall, Rabail Baig

Usefulness Workgroup: Keith Morse, Megan Salwei, Armando Bedoya, Bobby Barbarooah, Dennis Chornenky, Daniel Kortsch, Sawan Ruparel, Sadia Ali, Molly Beyer, Alejandro Muñoz del Rio, Morgan Hanger, Anthony Lin, Ashley Beecy, Ashutosh Singhal, Michael Phillips

Fairness Workgroup: Merage Ghane, David Talby, Mwisa Chisunka, Ted Robertson, Ariadna Vargas, Irene Dankwa-Mullan, Shawn Stapleton, Kellie Owens, Sana Khalid, Yiye Zhang, Kevin Larsen, Chevon Rariy, Allie Delonay, Anna Zink, Nadine Shillingford, Anietje Andy, Josh Karetny, Alaa Youssef

Safety Workgroup: Lee Fleisher, Nicoleta Economou, Chris Jackson, Aditya Kotecha, Tanvi Jayaraman, Tianqi Smith, Hou-Cheng Yang, Mark Kramer, Ilya Laufer, Anne Buckley, David Bates, Leigh Hanes, Brad Ulrich, Elizabeth Garwood, Rich Kenny, Suchi Saria, Tara Montgomery, Kevin Gormley, Aize Chao, Stan Huff

Transparency Workgroup: Lisa Lehmann, Shauna Overgaard, Agustina Saenz, Shyamal Sharma, Farhana Ferdousi Liza, Aaron Smith, Richard Moreno, Shannon Ridge, Anton Van Der Vegt, Christine Swisher, Sarah Newton, Ysabel Duron, Ted Gaubert, Robert Jenders, Brenton Hill, Joycee Berin, Paul Lukac

Security and Privacy Workgroup: Naomi Lefkovitz, Cora Han, Julie Snyder, Leila Murebee, Doug Williams, Gary Isaac, Deven McGraw, Alya Sulaiman, Richard Eng, Uttam Ghosh, Jennifer Brown

Independent Reviewers

Lisa Butler, Donna Cryer, Chethan Sabaru, Mozziyar Etemadi, Gary Weissman, Brett Moran, James Leo, Dom Pimenta, Subbu Venkataraman, Chris Mansi, Atul Grover, Bakul Patel, Timothy Hsu, Patrick Pavlick, Christopher Chen, Sven Zenker, Sabrina Hsueh, Jeffery Smith

Table of Contents

1. Preface

This Guide serves as a playbook for the development and deployment of AI in healthcare, providing actionable guidance on ethics and quality. It stems from the consensus-based approach of the CHAI community, drawing upon the collaborative work of patient advocates, technology developers, clinicians, data scientists, civil servants, bioethicists, and others. The Guide is written for an equally broad audience, encompassing everyone with rights and responsibilities in the process of designing, developing, deploying, and using AI technologies. To be both comprehensive and concise, the aims of this Guide are to integrate existing guidance and best practices into one cohesive framework, and to ground them in the real-world. A companion document, the Responsible AI Checklist (RAIC), elaborates the considerations found in this Guide at a finer level of detail, providing evaluation criteria for best practices across the AI lifecycle.

2. Summary

This Guide is written with multiple stakeholders in mind (see Section 3.2). It aims to foster a shared understanding among all parties on important considerations when selecting, developing and using AI solutions intended for patient care and related health system processes. The content of this Guide is organized around the health AI Lifecycle (see Section 4), providing considerations relevant to each stage in designing, developing and implementing AI solutions. Considerations have been organized by five principle-based themes at every stage of development and deployment:

1. Usefulness, Usability, and Efficacy
2. Fairness
3. Safety and Reliability
4. Transparency, Intelligibility, and Accountability
5. Security and Privacy

These core principles, expounded in Section 5, align with the National Academy of Medicine's (NAM's) AI Code of Conduct work [1], the White House Blueprint for an AI Bill of Rights [2], several frameworks from the National Institute of Standards and Technology (NIST) [3], [4], [5], as well as the Cybersecurity Framework from the Department of Health and Human Services Administration for Strategic Preparedness & Responses (HHS/ASPR) [6]. Harmonizing principles across documents, this Guide also translates principles from the abstract to the actionable, offering practical considerations for applying responsible AI guidance in day-to-day operational processes.

3. Introduction

For decades, healthcare has leveraged data-driven algorithms with a variety of applications and goals. With recent advancements in the field of AI, a new era of possibilities has emerged, enabling the capture and analysis of vast datasets thanks to increased storage and computing power. In broad terms, *AI* is a branch of computer science focused on developing techniques that enable computers to mimic intelligent behavior, akin to that of humans [7]. The term also applies to machine-based systems that can make predictions, recommendations, or decisions, thereby influencing real or virtual environments [8]. In high-stakes arenas like healthcare, it is generally preferable, and for many applications a requirement, to keep a human involved in the decision-making process with support from an AI solution. *Health AI* can be defined as the application of algorithmic systems to a suite of tasks including decision support, diagnosis, treatment planning, medical imaging analysis, patient monitoring, clinical note taking, precision medicine, and various administrative processes such as report writing, transcription of voice dictation, and summarization of text.

This Guide will not exhaustively define the various types of AI computing processes, which can range from rudimentary algorithms to traditional supervised machine learning to neural network models, including neural models that are referred to as *deep learning* and *large language models* (LLMs). Instead, this Guide presumes a general understanding of AI, focusing on it as a transformative healthcare technology. Health AI encompasses traditional machine learning, deep learning, and the multiple capabilities of AI systems (many of which have been supercharged in recent years by advances in machine learning), including natural language processing, computer vision, and other techniques that can augment medical expertise, improve diagnostic accuracy, streamline workflows, decrease workloads, personalize patient care, and enhance access and outcomes. This Guide is written with all such use cases in mind, including recent innovations in generative AI (see the Generative AI Use Case in Appendix 1).

Alongside many important demonstrations of AI's effectiveness, the risks of AI in healthcare are have been well documented, encompassing concerns about data privacy, biased or inaccurate results, the non-transparency of AI models, model drift, and workflow misalignment [2], [9], [10], [11], [12], [13], [14]. Accuracy issues can lead to incorrect diagnoses or treatment recommendations, while unexpected results and non-transparent models make it difficult for clinicians to trust and validate AI outputs. Model drift, in which AI performance degrades over time, further complicates effective implementation, as does the problem of workflow misalignment, where the AI solution does not integrate well into clinical processes.

Among these risks, bias stands out as one of the most high-profile problems, potentially resulting in unbalanced outcomes across different patient demographics. Although AI holds the potential to make healthcare more inclusive and effective, it also carries the risk of entrenching old patterns of bias and discrimination. Historically, the field of medicine has sometimes perpetuated social inequities along gendered, racial and economic lines, resulting in substandard care for underserved and underrepresented populations. Such disparities, exacerbated by social determinants of health, have led to adverse and traumatic outcomes for those affected. In some important cases, AI can help underserved communities, for example by activating care teams to

facilitate transfers to specialized care centers. But without shared standards of practice, rapid advancements in AI are still likely to worsen existing disparities, deepening the divide in accessibility and outcomes. Although AI practices are evolving around these concerns, they still lack consistent guidance on accountability, posing a risk to communities that have suffered most from harmful social determinants.

The last decade has seen increasing interest in the field of *responsible AI*, which considers such important factors as safety, reliability, fairness, and the ethical implications of AI systems and their uses. The work has been valuable as well as prolific, providing guidance to organizations on best practices and regulations. To date, more than 200 sets of AI guidelines have been issued worldwide by national governments and prominent organizations [15]. However, despite the proliferation of such documents, most of them remain somewhat abstract, never translating principles into the routine practices of everyday work. Furthermore, while each of these documents' advance similar themes and principles, there is no formal consensus among the myriad stakeholders involved. Different groups (e.g., developers, implementers, users, and regulators) may apply different ethics and quality frameworks, resulting in a fragmented landscape where responsibilities are not clearly defined or understood. The current lack of widely agreed upon guidance and best practices highlights a critical need for actionable guidelines that are broadly recognized and continuously updated. This ensures that all parties have a shared understanding of their obligations in the AI ecosystem.

In the domain of healthcare, the abundance of standards poses a challenge for technology developers who want to maintain accountability while meeting the needs of diverse health systems. Health systems, in turn, may face a dilemma when procuring AI solutions, lacking a transparent context for well-informed, safe, and beneficial acquisitions. Moreover, when guidance is opaque and poorly understood, problems with safety and efficacy may lead patients and clinicians to appropriately distrust AI recommendations. The situation highlights the need for all stakeholders to speak the same language, to follow the same or similar quality guidance, and to share similar delineation of responsibilities. In the absence of a shared understanding, the risks associated with AI multiply, compromising its benefits across different domains and workflows.

As stated above, the opportunities for AI applications have expanded so rapidly that many stakeholders, interested in formulating guidelines and guardrails for health AI, have done so in a decentralized manner. The situation no doubt reflects a shared feeling of both urgency and promise, but the undesirable consequence is a cluttered landscape of guidance, which creates challenges in both implementation and interoperability.

The Responsible AI Guide is a step toward greater collaboration and alignment, created to provide a cohesive approach that can extend across the landscape of health AI. To that end, this Guide reflects a multi-stakeholder effort to bring together multiple recommendations, guidance, and best practices now in circulation. By translating core principles into considerations, and by anchoring those considerations to real-world use cases, this Guide takes a concrete approach, bridging the gap between guidance and practice. Furthermore, it elaborates best practices for anyone involved in health AI, as laid out in Table 1 below.

To promote safety, efficacy, fairness, and trust in health AI solutions, and to cultivate a shared understanding of trustworthy AI for the health sector, the CHAI community introduced the "Blueprint for Trustworthy AI in Healthcare" [11]. The Blueprint lays out a plan for a comprehensive responsible AI framework, establishing a set of shared core principles for anyone developing and deploying AI. Building on the Blueprint, this Guide represents the next crucial phase of the framework, aiming to realize the benefits of AI while actively combating risks to usability, safety, fairness, and security. By offering tangible considerations to all accountable parties in the health ecosystem, this Guide ensures that the implementation of AI will be fair, transparent, safe, and useful.

The Responsible AI Guide and its companion, the Responsible AI Checklist (RAIC), are the result of a year-long effort conducted by CHAI workgroups. Beginning in April 2023, five workgroups convened weekly, each of them focusing on a subset of core principles described in the Blueprint, to craft considerations and evaluation criteria for stakeholders engaged in developing and implementing health AI solutions. Workgroups were composed of clinicians, data scientists, bioinformaticists, ethicists, patient advocates, civil servants, and people working at large and small technology development firms. The recruitment process accounted for gender and ethnic diversity, including faculty members from Historically Black Colleges and Universities. Through subsequent iterations and surveys to the broader CHAI membership, workgroups elicited stakeholder feedback at every step, aspiring to articulate consensus-driven guidance and best practices that would ensure widespread adoption.

Beginning with the development or procurement of an AI solution, this Guide follows the lifecycle of health AI through testing and deployment, raising considerations for the multi-disciplinary stakeholders involved (see Section 4). And, beyond merely raising considerations, this Guide makes operational recommendations at each phase of the AI lifecycle. The intended audience of this document thus ranges from technology developers to clinicians, from data scientists to bioethicists, from payers to policymakers, and from patients to caregivers (see Table 1).

When translating the considerations in this Guide to the real world, variations are expected within the context of each use case. For that reason, this Guide describes six example use cases to demonstrate such variations in considerations and best practices:

1. Predictive EHR Risk Use Case (Pediatric Asthma Exacerbation)
2. Imaging Diagnostic Use Case (Mammography)
3. Generative AI Use Case (EHR Query and Extraction)
4. Claims-Based Outpatient Use Case (Care Management)
5. Clinical Ops & Administration Use Case (Prior Authorization with Medical Coding)
6. Genomics Use Case (Precision Oncology with Genomic Markers)

While these use cases (further described in Appendix 1) represent only a sample of applications of AI in the health ecosystem, each of them relates to a broad family of use cases, offering a "paradigm case" so that readers can infer considerations for other practical scenarios (e.g., an EHR risk model might inform considerations for AI-driven clinical care at home). Given that these use cases are embedded in clinical systems, the scope of this Guide does not extend as far

as direct-to-consumer health AI solutions. Future iterations of this Guide, as it is used and adopted, will prove necessary as the field of AI develops. This first iteration of the Guide is best understood as a supporting structure for future responsible AI work, which will continue to evolve with the state of the field.

3.1 Scope Limitations of the Guide

The intent of this Guide is to offer actionable guidance in the domain of AI in healthcare. However, it cannot encompass all potential considerations or applications related to that domain. To ensure clarity on the focus of the Guide, the following limitations are also observed:

Focus on Healthcare Delivery: The discussion within this Guide centers on AI applications that directly impact healthcare delivery processes and patient care. Use cases addressing broader health-related areas, such as AI analytics for environmental factors like air quality, are not within the scope of this document. Future CHAI efforts will address other health related best practices on devices, public health, life sciences, and other sectors.

Geographic Context: This Guide is primarily embedded within the U.S. healthcare context, leveraging insights from EU documents and input from international members of the CHAI community. While efforts are made to ensure relevance across borders, nuances of specific regulatory frameworks outside the U.S. may not be fully addressed.

Environmental Impact: Despite the growing awareness around environmental sustainability in AI practices, this Guide does not address the environmental impact of AI systems utilized in healthcare settings. The environmental lifecycle of AI solutions, beginning with resources like rare earth minerals and water, requires consideration beyond the scope of this Guide. These complex and pressing issues may be subject to future consideration and exploration by CHAI.

Drug Development: While it is important to acknowledge the growing role of AI solutions in drug development, this Guide does not delve into that area. Future iterations and supplementary guidance from CHAI may provide more comprehensive coverage on the use of AI in pharmaceutical research and development.

Health Plan Implementation: This Guide focuses primarily on use cases within healthcare delivery. While it profiles one payer-focused use case (Prior Authorization, see Appendix 1), a more detailed exploration of AI implementation by health plans will require further work beyond the primary scope of this document. As with drug development, future guidance from CHAI may elaborate considerations in this area.

3.2 Stakeholder Types, Roles, and Representative Organizations

This Guide was developed in collaboration with patient advocates, and it is written with patient awareness in mind, acknowledging the pivotal role that patients play in the health AI ecosystem. While the technical language employed in this Guide may present some challenges for patient communities, efforts have been made to contribute to patient awareness throughout, focusing on the principle of transparency and intelligibility. The Guide serves as an initial step toward further collaboration between technical experts and patient communities, paving the way for more inclusive and patient-centered health AI practices in the future.

The table of stakeholders below represents the Guide's intended audience. While not exhaustive, it provides an overview of key participants in the health AI ecosystem, and it is sure to expand as AI technologies advance. Not all stakeholders may be formally considered "responsible" parties (see Section 3.3); however, stakeholders collectively bear responsibility to ensure that AI solutions are safe, fair, and effective. Their collaborative efforts are fundamental in shaping AI solutions that align with the principles at the core of this Guide (see Section 5).

In the Guide and across the AI Lifecycle (see Section 4 and Appendix 2), the *Developer Team* refers to stakeholders primarily involved in the AI solution development process and the maintenance of the solution; they may consist of data scientists, software engineers, data engineers, user experience and interface designers, or a subset of those. The *Implementer Team* comprises stakeholders involved in implementing, using and integrating an AI solution in health system workflows, including but not limited to healthcare providers, human factors and behavioral science professionals, health system leadership, health system information technology professionals, etc.

The stakeholders listed in Table 1 also play a crucial role by establishing the frameworks and policies that guide ethical and high-quality AI development and use. Their involvement ensures that AI solutions adhere to standard that prioritize usefulness, fairness, safety, transparency, privacy, and security. Through their governance, stakeholders help foster trust and accountability in the AI solution, which is essential for their acceptance and integration into workflows.

The primary concern of any AI solution should be the communities impacted by the model, such as patients and their caregivers. Their well-being and outcomes are the driving force behind the responsible development and implementation of health AI solutions. The design of an AI solution should account for both the user's and patient's journey within its workflow, and evaluations of the AI solution should factor in the experience of both. When AI solutions are intended for direct patient use, it is essential to employ human factors design methods to comprehend both the capabilities and constraints of the end users. Overall, patients have rights to receive information about, benefit equally from, and influence oversight on AI solutions [16].

Table 1: Stakeholder Roles, Professions and Representative Organizations

Stakeholder Roles	Example Stakeholder Professions	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 Licensing Bodies, Medical Device Collaboratives, etc.

3.3 Driving Accountability for Safe, Effective and Responsible AI

The CHAI community aims to create guidance and best practices by providing detailed and context-specific considerations at each stage of the AI Lifecycle (see Section 4, Section 7, and Appendix 2). This includes specifying the accountability and responsibilities of developers and healthcare organizations, as well as identifying the roles of anyone responsible for specific actions. Organizations should identify who is accountable for what and to whom they are accountable, as well as who the end users of an AI solution should contact with issues or concerns, and who will take action in response.

Incorporating accountability into AI solutions will ensure that developers and implementers are held responsible for the ethical and effective deployment of AI in healthcare settings. It involves establishing mechanisms for oversight, governance, and quality control to monitor and evaluate the performance of AI systems over time. It also entails a scope of responsibility beyond existing regulations, which nevertheless coheres with the evolving landscape of ethics and standards.

Identifying roles and responsibilities within an organization facilitates accountability and change management, as this process gives responsible parties a common understanding of their responsibilities. Clear documentation of responsibilities in the context of processes and procedures is necessary to achieve AI solution quality and ethical use. Important roles in the context of responsibility and accountability within an organization include the *business owner*, *technology owner*, *executive sponsor*, and *end user* [17].

The *business owner*, who is typically a member of the implementer team, is the individual who articulates the need for the AI solution, helps in certain cases with its development, tests the AI solution for its performance and utility, and assesses its impact. This individual drives the adoption of the solution and serves as champion for its use, as well as acting as primary point of contact for addressing any risks when the AI solution is operating. In terms of the stakeholder roles in Table 1 above, a business owner may be an end user (such as a healthcare professional,

healthcare administration staff member) or clinical administration staff member. For clinical workflows, this individual is typically a licensed clinician or clinical staff member.

The *technology owner* is responsible for the technical aspects of the AI solution, including its functions and maintenance during deployment. In terms of stakeholder roles, technology owners will typically be a data science developer or informatics and information technology professional. At present, a typical health system may lack experts with necessary training in data science and AI, which is a pain point deserving attention. Technology Owners play a pivotal role as the conduit or point person ensuring adherence to best practice, and serve as the primary point of contact for addressing risks and driving technology adoption. Together, business and technology owners ensure that AI solutions are developed and implemented ethically and effectively.

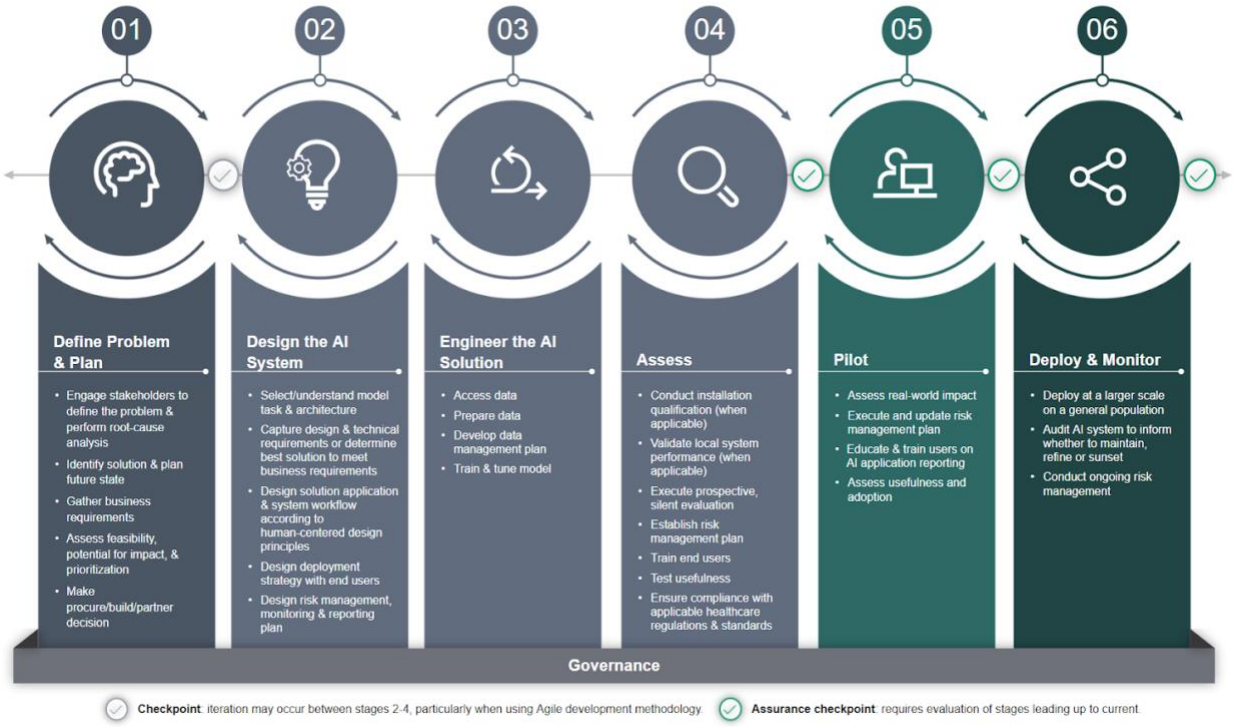
An individual from the implementer organization's leadership should be assigned to serve as an *executive sponsor*, aligning the implementer team and the AI solution with the organization's strategic priorities and resources. The *end user* is responsible for reporting risks and concerns through appropriate channels, as defined by the implementer organization when the AI solution is used for patient care.

Responsibility and accountability are crucial for considerations across the AI Lifecycle. To mention a few examples: When first defining the problem (Stage 1), the developer team’s organization is responsible for ensuring that the intended purpose of the AI technology meets the needs of both the implementer organizations and the market, based on the business requirements gathered. When designing the AI solution (Stage 2), the developer collaborates with the implementer to design the model based on the business requirements defined by the implementer. The implementer, in turn, designs the workflow to integrate the AI solution into clinical practice, addressing the specified needs. In the same stage, the implementer defines the impact measures (e.g. clinical outcome measures) to evaluate the AI solution’s effectiveness during testing and use. A risk management plan, detailing mitigation strategies and responsible parties, is drafted in the Design stage (Stage 2), updated during the Engineering phase (Stage 3) and again during the Assess stage (Stage 4), clarifying who will report safety risks or harm to the developer and implementer organizations. When engineering the AI solution (Stage 3), individuals and teams responsible for data monitoring are established along with stakeholders responsible for receiving their reports. During the Pilot and Deployment stages (Stages 5 and 6), responsible parties are identified for tracking adverse events, determining follow-up actions, and communicating information to affected stakeholders. Lastly, governance processes span the lifecycle, and they help identify who is accountable for reporting on AI solution development, performance, issues during deployment, and impact across the implementer and developer organizations.

4. The AI Lifecycle

Through a consensus driven process, CHAI has developed a 6-stage health AI lifecycle (see Figure 1) based on industry-standard AI development frameworks [11], [18], [19], [20], [21], augmented with implementation considerations and responsible AI recommendations critical to healthcare settings. The lifecycle encompasses processes from initial problem identification and solution planning through to large-scale deployment and monitoring of an AI solution and surrounding system (workflows, technical support components, personnel). New AI methodologies and solutions may be developed independently of the lifecycle described herein. However, before advancing an AI solution to deployment, the developer team should collaborate with implementers to ensure that the AI solution meets the needs of the use case and adheres to best practices for trustworthy AI. Additionally, developer teams may conduct robustness testing of the AI model to evaluate its performance and scalability across different patient populations.

Figure 1: The CHAI 6-Stage Lifecycle for Health AI Development and Deployment



The CHAI 6-stage AI lifecycle spans processes to (1) Define the Problem & Plan, (2) Design the AI System, (3) Engineer the AI Solution, (4) Assess the System, (5) Pilot the System, and (6) Deploy & Monitor the System. These stages are not always linear, and there may be feedback loops between them. For example, the results of an assessment in stage 4 may lead to changes in the design or model training in stages 2 or 3 respectively.

Key to the lifecycle are a series of checkpoints (after Stages 1, 4, 5 and 6). These checkpoints ensure that AI solutions are independently reviewed (see Section 6) and plans are in place for ongoing high quality patient care and ethical use of the AI solution. Responsible AI checkpoints are crucial, both before and after an AI system is in use (during silent deployment and pilot stages, then again at regular intervals during general deployment). These checkpoints have a significant impact on health system processes and patient care, ensuring safety and security during active system usage. Each checkpoint incorporates measures like key performance indicators, model performance metrics, user satisfaction assessments, and ongoing monitoring of output.

Governance throughout the lifecycle ensures that organizations are well-prepared to conduct responsible AI assessments at each stage, and is represented in Figure 1 as an underlying foundation. The CHAI AI Lifecycle forms the basis for a series of *core principles and considerations* (presented in detail in Section 7) that health systems and AI solution developers should take into account when developing, procuring and assessing AI system deployments.

In the subsections below, each lifecycle stage is briefly described, along with relevant decision points for determining when it is complete. As visually depicted in Figure 1, completion of a stage may be followed by a return to an earlier stage to add or modify relevant components. Detailed descriptions of each stage are provided in Appendix 2.

Stage 1: Define Problem and Plan

Healthcare is riddled with "solutions" in search of problems. Responsible innovation requires a clear understanding of the specific issue AI is intended to solve, which will drive the intended purpose of the tool. The intended purpose a) defines the scope of verification & validation activities and b) allows reasonable delineation between user responsibility and vendor responsibility. First, however, an upfront investment of time and effort is needed to map root causes and understand the specific needs of those experiencing the problem(s). Stage 1 focuses on this process, with developer teams conducting surveys, interviews or market research to understand healthcare system needs. In Stage 1, implementer teams also identify a clear problem, its setting, and stakeholders involved (which together comprise a *use case*), thereby quantifying potential return on investment, return on health outcomes, and improvement of healthcare operations. Implementers use this information to decide whether to build an in-house AI solution, procure one from a third party, or partner with a third party to develop a solution. Stage 1 consists of 5 steps: (1) Engage stakeholders to define the problem and perform root-cause analysis; (2) Identify solution and plan future state; (3) Gather business requirements; (4) Assess feasibility, potential for impact, and prioritization; (5) Make procure/build/partner decision. At the end of Stage 1, after defining the problem, the implementer team faces a key decision: whether to procure, build, or partner to develop the solution.

Stage 2: Design the AI system

Stage 2 involves capturing details of the AI system of which the solution is the main component. This includes detailing its technical requirements, proposed system workflow, and deployment strategy. The design of the AI system is informed by the business requirements of the health system(s) and the needs of representative end users where the AI solution will be implemented. Stage 2 has 5 steps: (1) Select/understand model task and architecture; (2) Capture design, data, technical requirements, and scope, or determine the best solution to meet business requirements; (3) Design solution application and system workflow; (4) Design deployment strategy with end users; (5) Design monitoring and reporting plan. The resulting designs will be used as the basis for engineering the AI solution (when applicable) or to determine with the developer team how a current commercially available solution should be adapted. The implementer team will determine the strategy for how the AI solution should be deployed within the workflow.

Stage 3: Engineer the AI Solution

The engineering stage aims to create an AI solution that can accurately predict or classify data and develop the interface for exposing AI solution output, as defined during the Design stage. This stage also ensures that AI solution deployment can be operationalized and that adequate planning is completed prior to deployment. In cases of externally developed AI solutions, the developer should provide expertise in collaboration with the implementer, who ensures that the AI solution meets its intended purpose via risk-benefit analysis before and after deployment. Stage 3 has 4 steps: (1) Access data; (2) Prepare data; (3); Develop data management plan; (4) Train and tune the model underlying the AI solution to meet its intended purpose. This stage culminates in a quality-assured dataset with documentation supporting lineage, and a fully-developed model with validated outputs and, where possible, impact. (In certain instances, the impact of the AI solution can only be assessed only in a real-world setting.) With the data and model in hand, the team may advance to the next stage for a business decision of whether to deploy the AI solution into the healthcare system, or, when applicable, return to the Stage II to refine the design of the AI solution or corresponding workflow.

Stage 4: Assess

This stage involves a series of assessments to determine whether to proceed with a pilot of the AI system in Stage 5. When existing AI technologies are acquired from an external organization, local validation and installation qualification need to be conducted first, prior to the assessment of the AI system as a whole. A change management plan should be in place to delineate who, between the developer and implementer, is responsible for performing these duties. This is followed by a prospective, silent evaluation and the establishment of a risk management plan, based on anticipated risks from Stages 2 through 4. These steps are followed by end user training and usefulness testing, along with a review to ensure compliance with applicable healthcare standards and regulations prior to piloting and deployment. Stage 4 consists of 7 steps: (1) Conduct installation qualification (when applicable); (2) Validate local system performance (when applicable); (3) Execute prospective, silent evaluation; (4) Establish risk management plan; (5) Train end users; (6) Test usefulness; (7) Ensure compliance with applicable healthcare regulations and standards. A business/clinical owner should be defined, who will be accountable

for ensuring that the AI solution is tested and that personnel are trained, eliciting their feedback. This stage culminates in a business decision to deploy the AI application (or not) as a pilot. The decision to pilot is accompanied by approved implementation, measurement, and mitigation plans, as well as pilot user training, prior to deployment.

Stage 5: Pilot

The pilot stage is the first real-world use of the AI solution by the implementer team to inform large-scale deployment plans. Prior to the general deployment of an AI system, careful review and consideration must be made by the health system to decide whether or not to deploy an AI model into production. Based on this pilot stage, success criteria are reviewed to inform the decision on whether to deploy the AI system. Some common success criteria include the AI solution's accuracy, reliability, interpretability, feasibility, user acceptance, cost, and alignment with the organization's values and goals. This process is primarily undertaken by the implementer team. The pilot stage has 4 steps: (1) Assess real-world impact; (2) Execute and update risk management plan; (3) Educate and train users on the AI application, its intended purpose and use, and reporting. The decision point at the conclusion of the pilot is whether to proceed with a larger scale deployment; (4) Assess usefulness and adoption, evaluating workflow integration, end user acceptance, and potential downstream impacts of the AI solution.

Stage 6: Deploy and Monitor

The deployment and monitoring stage is the process of making the AI solution and system broadly available to the healthcare system or relevant specialty. Once deployed by the implementer team, the AI solution is often handed over to a model operations team (when available) to provide ongoing monitoring, retraining, and governance of models to ensure peak performance and that decisions are transparent. This stage has 3 steps: (1) Deploy at a larger scale on a general population; (2) Audit AI system to inform whether to maintain, refine or sunset; (3) Conduct ongoing risk management. This stage culminates in a successfully deployed AI system with ongoing monitoring. If and when AI solution performance drifts or deviates, the AI solution may be revised, possibly returning to Stage II or Stage III, or the AI system may be decommissioned entirely.

5. Core Principles for Trustworthy Health AI

The Responsible AI Guide is built on a set of core principles for trustworthy health AI. From these core principles and their related concepts, CHAI has developed a set of *considerations* relevant to each stage of the AI lifecycle. These considerations encompass factors that should be assessed when considering strategies for designing and deploying an AI solution for use in a healthcare system (see Section 7).

1. **Usefulness, Usability, and Efficacy.** To be **useful**, an AI solution must provide a specific benefit to patients and/or healthcare delivery, and it must prove to be not only

valid and reliable but usable and effective [11]. The **benefit** of an AI solution can be measured based on its effectiveness in achieving intended outcomes, as well as its impact on overall health resulting from both intended and potentially unintended uses. An assessment of benefit should consider the balance between positive effects and adverse effects or risks [11]. The usefulness of an AI solution also depends on the cost of its deployment and the capacity of personnel to take action as a result of its output or guidance [22]. Relatedly, an **effective** AI solution can be shown to achieve the intended improvement on health compared to existing standards of care, or it can improve existing workflows and processes; for example, an AI solution intended to increase the efficiency of a workflow can be associated with reduced costs or shorter times to complete tasks [1], [23].

The **robustness** of an AI system can be demonstrated by its ability to maintain its level of performance under a variety of circumstances [24]. The solution's **testability** is more encompassing, demonstrating the extent to which its performance can be verified as meeting all principles for trustworthy AI including safety, fairness, transparency, privacy, and security [11].

The **usability** of an AI solution connotes the quality of the user's experience, including effectiveness, efficiency, and satisfaction with the technology [11], [25]. In this context, an **engaged** human-centered development and deployment process entails understanding, expressing and prioritizing the needs, preferences and goals of end users and other stakeholders, as well as considering related implications throughout the AI lifecycle [1]. Similarly, an **accessible** development and deployment process ensures that stakeholder access and engagement is a core feature of each stage of the AI lifecycle and governance [1]. Additionally, an **adaptive** accountability framework ensures continuous learning and improvement, providing ongoing information on the results of the AI solution [1], [14].

2. **Fairness**. To be considered **fair**, AI solutions require (1) **parity**, meaning that common measures of algorithmic performance are equal across protected subgroups; (2) **calibration**, meaning that outcomes are independent of protected characteristics (or class) – such as race, gender, or their proxies; and (3) **anti-classification**, meaning that protected characteristics are not explicitly used to make decisions. Fairness applies beyond diagnostic accuracy to balanced allocation of resources, access to care, and outcomes [26], [27]. Following from that, the ultimate indicator of fairness goes beyond these measures and should be accompanied by measures showing (4) comparability in access to care, outcomes, and resource allocation. It is important to combat disparities in health (or in the major social determinants of health) between groups with different levels of underlying social advantage or disadvantage – that is, wealth, power, or prestige [28].

   Under the headings of fairness, the risk of **bias** is multifaceted and warrants a taxonomy of definitions. **General bias** is a distortion towards a particular perspective, outcome, or interpretation that can result from systemic, social, emotional, operational, or statistical tendencies and limitations. **Population bias** occurs when AI solutions duplicate social stereotypes, particularly toward protected groups – based on gender, age, race, religion, social status, or others – in a way that leads to reasoning errors [27], [29]. **Data bias**

arises when there is skew in data collection (measured, unmeasured, unmeasurable, and more), sampling, transcriptions, or observer/interpretation which can lead to a limited and biased reflection of facts [30]. **Algorithmic or model bias** develops when a machine learning (ML) algorithm produces results that are systematically prejudiced due to improper feature selection or engineering, flawed assumptions in the algorithm's design, and improper training approaches (e.g. biased cohort selection for case/controls, inadequate feature selection, improper feature engineering, improper imputation, target leakage, model architecture, training hyperparameters, loss-function, regularization, etc.), all of which can lead to unfair or discriminatory outcomes [27], [31]. **Interpretation or use bias** arises from human behaviors, perceptions, and interpretations based on their experience rather than from the AI solution itself, as well as how their experience and understanding of the data influences interpretation of model predictions [27]. In a similar vein, **automation bias** occurs when a user defaults to the recommendations of a model without integration of additional (but necessary) information [27], [32]. This may result from time pressures, low energy levels, or workflow constraints. These bias types are not exhaustive, especially given the range of social, emotional, and cognitive biases that sway human decisions and behavior, but they are the most common types of bias currently subject to evaluation.

3. **Safety and Reliability**. A **safe** AI solution does not endanger human life, health, property, or the environment. In healthcare, this translates to the avoidance, prevention, and amelioration of AI-related adverse outcomes affecting patients, clinicians, and health systems [11], [33], [34]. Harms or a diminishment of safety may occur due to misuse or model deterioration because of factors like drifts and shifts [14], [35]. An AI system therefore proves **reliable** to the extent that it can perform as required without failure, incorporating backup plans that ensure continuity, resilience, communication, accountability, and responsive action in the event of any issues [24].

4. **Transparency, Intelligibility, and Accountability**. In this context, **transparency** is the extent to which information about an AI solution (e.g., capabilities, limitations, and purpose) and its output is available to all relevant stakeholders [5]. **Intelligibility** is the extent to which the AI system can be understood by relevant stakeholders, often through a representation of the mechanisms underlying an algorithm’s operation and through the meaning of its output in the context of its designed functional purposes [36]. The principle of intelligibility addresses the question of whether humans can understand and make sense of the AI solution as a whole, encompassing the principles of explainability and interpretability. **Explainability** is the ability to provide insight into why and how the AI model is generating outputs – the observation of the inner mechanics of the AI/ML method, along with the factors and features that influence the system’s decision-making process [5]. Explainability addresses the question of why an AI system made a specific decision. **Interpretability**, by contrast, is the ability to understand the cause and effect of the AI model’s output in human terms, serving as a risk mitigation strategy. It involves making the model structure, parameters, and relationship between inputs and outputs understandable [5]. **Observability**, then, connotes the ability to observe inputs, outputs, impact, and consequences of model predictions. In applications designed to augment human decisions, a **human-machine teaming** model should explicitly specify the

interface for interaction between the AI solution and the human user [23]. This specification lays out the details needed for the user to operate the model safely, which finally supports the principle of **accountability**: the responsibility and liability for minimizing harm throughout all stages of the AI lifecycle [24], [37].

5. **Security and Privacy**. The principle of **security** conveys the extent to which AI systems can maintain confidentiality, integrity, and availability through administrative, technical, and physical safeguards [5]. **Privacy** is the extent to which AI systems can maintain predictability, manageability, and dissociability through administrative, technical, and physical safeguards that prevent problematic data actions for individuals (including at the group and societal level) [4].

   In this context, **risk** is the composite measure of an event's probability of occurring and the magnitude or degree of consequences resulting from the corresponding event [5]. **Risk management** is a term that signifies coordinated activities to direct and control an organization with regard to risk [5].

5.1 Core Principles in Context

In the realm of healthcare, ethics and quality principles are translated into clinical practice through the implementation of standards [38]. Throughout the AI lifecycle, organizations can evaluate AI solutions by gathering evidence that they align with relevant principles, conducting thorough testing and local validation to assess the risks and benefits of each AI solution, evaluating its impact on health or healthcare delivery. Randomized controlled trials and other well-designed research methods are crucial for building an evidence base to ascertain the effectiveness of AI tools in clinical settings.

Best practices should, however, be tailored to each use case, since risks and benefits may vary depending on the context of a given AI solution and its intended purpose. For instance, an AI solution intended for diagnosing life-critical events, such as AI solutions that aid the detection of breast cancer, may prioritize safety considerations more heavily than those designed to aid end users with administrative tasks. Numerous factors play a significant role in determining the potential risks and benefits of an AI solution, like the context of the patient population, the severity of the healthcare situation (be it critical, serious, or non-serious), and the significance of the healthcare decision within the workflow [39]. Moreover, when the end user of an AI solution lacks the necessary education or background to understand the output, accuracy, risks, and limitations of the AI output, the situation presents another risk to be considered, measured, and managed appropriately.

For algorithms that meet FDA criteria for software as a medical device (SaMD), a risk-benefit assessment is part of the standard evaluation [40]. But in broader terms, the risks and benefits of an AI solution can be weighed differently by each organization when making decisions about the AI solution's potential deployment and use. According to NIST, risk tolerance is the level of risk

an entity is willing to assume in order to achieve a potential desired result [41]. Organizations may be willing to accept more risk when high benefit is demonstrated in certain use cases, similar to cancer therapeutic trials aimed to address the high mortality rate of conventional therapies. Risk tolerances may vary across organizations depending on their different business environments, culture, and core values.

5.2 The Role of Governance

Within each organization, AI governance can operationalize and implement standard requirements translated from quality and ethical principles. Beyond local oversight, governance structures can ensure that an independent review is performed for AI solutions, thereby meeting ethical requirements and high-quality outcomes. AI governance structures can drive accountability through SOPs and by specifying roles and responsibilities.

Responsibility for the oversight of areas like safety and security should rest high enough in the organization so that decisions can be made promptly about resources, risk mitigation, incident response, and potential rollback of AI systems [39], [42], [43]. Along the same lines, organizational stakeholders should establish and maintain clear governance policies and procedures to manage risks and changes, taking into account organizational mission priorities, risk tolerances, and legal and contractual obligations [3], [4]. As well as managing risks to safety and security, organizations should define a bias management structure capable of evaluating an AI solution's fairness across the AI lifecycle [44], [45], [46], [47]. This entails a clear layout of roles and accountability, including all stakeholders and anyone responsible for independent evaluations or audits (see Section 6 below) [39], [42], [43]. Organizations should ensure that anyone involved in the selection, development, and deployment of AI solutions is well-trained on principles and considerations in ethics and quality [48]. In addition, relevant personnel should receive training on processes like safety reporting, and they should be made aware of change management agreements and processes. Workforce training should simultaneously include information about the intended use, risks, limitations and implications of AI solutions [3], [4], [40]. Risks and limitations ought to be framed by organizational standards for "adverse event" (AE) and "serious adverse event" (SAE) so that impacts and risks can be assessed accordingly [49]. Lastly, organizations should establish a clear approach for handling transparency and AI intelligibility to the public, enhancing awareness and fostering trust [49], [50].

6. Importance of Independent Review

When AI systems are deployed in high risk environments like healthcare, independent quality evaluation is crucial to ensuring patient safety, efficacy, and trust in the underlying AI solution. Commercial companies with potential conflicts of interest may develop AI solutions; however, even without considering the factor of commercial influence, external scrutiny can help uncover unintended and unforeseen risks. A rigorous and standardized process of independent review can provide objectivity when identifying technical flaws, biases, or unintended behaviors that AI

developers may have missed or minimized, the consequence of which would adversely affect patients and society [51], [52]. Uncovering issues prior to deployment in healthcare environments will allow developers and healthcare systems to mitigate risks, and ensure the safety, reliability, and clinical utility of AI systems. Independent review can also promote transparency and accountability. It can also promote sharing information about both successes and failures among stakeholders to foster industry growth. Setting a standard where the findings of independent reviews are publicly accessible can build public trust in AI systems and hold developers accountable for the output of an AI model.

Healthcare has a strong history of quality control and independent review standards for medications, devices, and lab assays. In the US these are governed by the agencies within the Department of Health and Human Services, where a comprehensive process of external review protects the public's health and safety. The Food and Drug Administration sets regulatory standards for medications and devices, ensures safety and efficacy through the review of clinical trial data, and provides independent oversight to counteract potential biases or conflicts of interest [53], [54]. This process maintains the confidence of patients and clinicians, and it also facilitates alignment with regulatory authorities throughout the world.

To achieve the goals of independent quality evaluation, AI solution developers need to adopt clear standards and benchmarks for evaluating AI solutions on measures of safety, reliability, bias, fairness, and efficacy. The adoption of standards and benchmarks may facilitate an initial internal organizational review, which could then be validated by an external certified quality assurance laboratory. Independent quality evaluation should also ensure transparency regarding data used to develop models, AI methods, and validation of models, as well as risks and limitations. Such practices are essential to accountability and public trust in AI systems. As many algorithms continuously learn and are updated, it is also important to establish a process of iterative review and data sharing, similar to the process implemented by the FDA for post-approval monitoring and adverse event reporting. Again, such practices are essential, because they facilitate continuous learning and evaluation, supporting goals for the long-term safety and efficacy of AI systems.

7. Implementing Best Practices for Trustworthy Health AI

This section of the Guide is its centerpiece, encompassing its most substantial content. Elaborating on considerations from each core principle described above, it discusses how to implement those principles in the form of assessments at each stage of the AI Lifecycle, while also demonstrating use case-dependent variations at the end of each stage. Considerations may be repeated with nuances at different stages or under different principles. The present iteration of the Guide includes all such repetitions to provide sufficient coverage. However, the Responsible AI Checklist (RAIC) consolidates evaluation criteria stemming from all such considerations, and

it is designed for independent review or assessment by parties involved in development and deployment.

Stage 1: Define Problem & Plan

Usefulness, Usability, and Efficacy: Stage 1

Clearly define the problem to be solved and explain why the AI solution is necessary [55], [56], [57], [58]. This includes considering whether a given AI solution will address the stated use case, is consistent with organizational objectives, and whether it could potentially improve on the standard of care or existing practice.

Consider how the AI solution will integrate into the workflow [58], [59], [60], [61]. This requires completing and documenting a workflow integration assessment, which includes accounting for how the AI solution will affect flow of people and tasks in both physical and digital environments. Importantly, it also includes assessing the potential for impact on patient-clinician interactions.

Assess benefits, risks, and costs associated with deploying the AI solution [12], [62]. A systematic approach for evaluating relative risks and benefits (including a cost-benefit analysis) of deploying an AI solution is helpful for this step, which requires that potential benefits and risks be identified and documented.

Evaluate whether end users are likely to trust the AI solution and its output [63], [64], [65]. This step includes determining whether the solution's functions are transparent and understandable to end users and its limitations are explained in non-technical terms, including scenarios in which the solution is not expected to perform well for a given use case. The potential impact of the solution's risks and benefits on user confidence should also be assessed, and a pathway should be created to address end users' concerns.

Ensure that relevant clinical experts have been involved in the development and clinical validation of the AI solution [12], [66]. As part of this step, assess whether the clinical validation success rate has been measured against medical criteria.

Fairness: Stage 1

Assess whether the framing of the problem addressed by the AI solution inherently disadvantages or discriminates against specific socio-demographic subgroups [27], [45]. Consider whether the problem definition and its associated solution are broad enough to cover diverse scenarios that are not restricted to a subset of the population.

Define and apply *fairness and balanced access, outcomes, and resources* in the context of the problem and its AI solution [67], [68]. Assess whether the implementer team has defined how fairness will be evaluated in the context of the AI's performance for the use case, and for whom. In addition, ask whether this definition includes concepts of minimizing harm and maximizing both clinical access and benefit.

Establish a bias monitoring and mitigation strategy [27], [44]. The team should determine whether there are differences in feasibility or effectiveness for relevant subgroups or end users based on workflow (e.g., language limitations, access limitations, insurance limitations, provider limitations, provider patient load, etc.). They should also determine whether there are security safeguards in place to protect against actions including intentional data contamination and model-based attacks.

Identify relevant socio-demographic subgroups that may be at risk of bias [9], [68]. The team should evaluate the AI solution for the potential to amplify existing social inequalities, and whether it is possible that *not* considering socio-demographic subgroups could cause harm (at individual or population level) or reduce overall generalizability. The team should also determine if there are documented criteria for ensuring AI fairness across all subgroups.

Identify potential types and sources of AI deployment bias [27], [69]. These include workflow or data variability that could contribute to bias after deployment. If the AI solution will be deployed in multiple settings, consider whether the patient population varies across settings and whether there is a formal plan to test for those biases across sites. The team should also ascertain whether there are systematic differences between the solution's training data source environment and deployment context in terms of workflow, treatment protocols, provider types, patient load, population representativeness, accessibility, data sources, and IT service integration.

Identify when and how users and impacted populations can provide feedback related to fairness and bias in the design/workflow of the AI solution [49]. This includes determining whether mechanisms are in place for stakeholders and end users to provide feedback or raise issues regarding potential bias and fairness of operational processes.

Ascertain suitable methods for conducting bias risk assessment and management [27], [70]. Teams should determine whether methods used to assess and manage the risk of bias, especially concerning relevant subgroups, are documented, including clear, predefined considerations or assumptions informing AI bias risk assessment for relevant subgroups. If such documentation exists, determine whether it includes a process for regularly updating the bias risk management framework.

Determine whether externally acquired AI solutions comply with privacy and data security policies [71]. The team should assess whether data use/sharing agreements align with policies regarding personally identifiable information (PII) and conform to HIPAA requirements.

Consider industry partner accountability through the fairness and bias evaluation process [12], [72]. For AI solutions procured from a vendor, determine whether the vendor can provide documentation of bias evaluation steps taken, metrics, and outcomes of those steps (based on the implementer team organization's AI/ML bias policies and relevant definitions of fairness). In addition, determine whether the vendor will share AI system performance and parity information according to relevant demographic subgroups and allow a third-party organization to conduct a bias evaluation to meet the implementing team organization's needs, policies, and guidelines. Teams should make sure to allocate sufficient time to conduct a performance evaluation and bias/fairness assessment.

Consider how patient and/or population data will be shared, and assess the potential impact of data sharing on fairness and bias [73]. The implementing team should determine whether patient data will be shared with a third-party vendor as part of an AI system purchasing agreement (for example, data passing through remote vendor servers to produce performance metrics). The team should then assess whether processes are in place to maintain privacy and safety of patient data consistent with data use agreements. Vendors should be asked whether they guarantee that data will not be used to predict sensitive health information or identity information outside the context of the AI solution.

Consider industry partner transparency regarding fairness and bias evaluation in model performance, parity, and balanced access, resources, or outcomes for relevant socio-demographic subgroups [73]. If an AI system is procured from a vendor, determine whether the vendor is able to provide clear, stepwise information on how the AI/ML system was developed, including sources of data used to train the underlying model, and who developed it. Assess whether the training data is representative of the deployment context.

Consider how third-party security practices can expose internal data to risk or bias [73]. The implementer team should consider whether their organization has evaluated how internal and vendor security practices could expose AI models or data to external attacks. The team should also determine whether there are practices in place to minimize the scope and degree of impact from such attacks, especially ones that could result in data theft or biased data distributions, or alter model attributes/function in ways that could expose specific subpopulations to greater risk of harm.

Apply fairness and bias considerations when determining the optimal balance of human judgment and AI-based decision-making [74], [75]. Important questions to ask at this point include: Will end users take information from *only* the AI solution to make decisions, or will they integrate its output with other information? Will they have the option to override an algorithmic decision, and are there clear guidelines around that option?

Safety and Reliability: Stage 1

Perform a current state analysis to identify potential harms and risks [5], [12], [39], [49], [76], [77]. A *current state analysis* includes feedback from representative end users during selection/design stages to inform risk management practices for the deployment of the AI system (see Stage 2). This may include elements such as the integration of human factors into harm assessment during definition of the use case, the return on health (ROH)/ return on investment (ROI) analysis, and selection of the AI solution, as well as determining whether safety, bias, security, and other risks have been identified by end users and others during design/planning. In addition, the team should compare the potential safety risks of the proposed solution with current-state safety data. They should also determine whether a risk management process exists to evaluate risks to both patients and users, including requirements and strategies for mitigating harm. Finally, the team should also consider whether the organization expects a risk management plan to be in place for each AI solution, articulating risks and potential issues and how those will be managed, and whether the organization has risk management SOPs in place to ensure consistent decision-making for identified risks.

For the AI solution being selected or developed, establish clear inclusion/exclusion criteria for the targeted patient population [49]. The team should determine whether a protocol exists for these criteria, but in certain cases populations are not under hard exclusion rules.

Ensure that the developer and the implementer organizations are responsible for the safety, effectiveness, and performance of the AI solution throughout its lifecycle [14], [35], [39]. Agreements around responsibilities should be established early in the selection process. For example, when an AI solution incorporates an off-the-shelf database system, the implementing organization should understand its capabilities and limitations throughout its lifecycle. The implementer team should also determine whether an agreement has been established early in the selection process regarding responsibilities of all involved parties. The team should also explore whether their organization is aware of limitations of the underlying technology and underlying data at the implementing site(s) and whether they are aware of any alternatives or modifications needed to ensure patient safety.

Conduct an initial assessment to ensure compliance with federal and local regulations [49]. Assessing an AI solution for regulatory compliance involves checking if the technology meets the FDA's criteria for Software as a Medical Device. The implementing team should confirm whether this assessment has been completed and if the technology falls under FDA oversight as outlined in the FDA's Digital Health Policy Navigator. Additionally, compliance with regulations from other agencies such as the ONC should be evaluated, along with local policies and procedures, including IRB guidelines as applicable.

Consider ethical and legal challenges and how they will be handled [49]. This step includes ascertaining whether legal considerations have been taken into account (e.g., what happens if an AI model is not FDA approved?), and whether there have been related cases or lawsuits that the implementer team or end users should be aware of. Identify if there are local laws (state) that

enforce safety reporting, or laws around informed consent that have implications for implementing the proposed AI solution. A further question to explore is whether patients will be made aware that AI is being used, especially in case of adverse events. In addition, determine whether there are mechanisms in place to disclose adverse events, and whether there is a protocol to ensure that researchers are informed of safety issues.

Transparency, Intelligibility, and Accountability: Stage 1

Document how the problem/solution justifies the use of AI [12], [78], [79], [80]. For transparency purposes, ensure that a clear argument is documented for the use of the tool, over and against alternatives. Assess if the medical context and rationale is appropriate, whether the workflow evaluation is complete and accurate, and determine which key performance indicators are necessary to measure the tool's impact.

Consider the purpose of the AI solution [12], [66], [80], [81], [82]. This includes determining whether the AI solution's intended use has been documented, including its intended users (e.g., healthcare professionals, patients, the public). It includes documenting plans for how and when the outcome will be assessed.

Consider the accessibility of project-related information and model-related information to project stakeholders [81], [83]. This step includes choosing the format (e.g., Model Card) in which to communicate information about the AI solution to project stakeholders, developers, end users, and patients. Additional considerations include whether end users and patients will have access to the same documentation as stakeholders involved in technical aspects of implementation and piloting, and whether those decisions adequately account for the needs of both patients and end users.

Determine what types of information should be documented [81]. This documentation may include an overview of the AI solution (i.e., who is developing it, date, version, type, citation details, license, etc.), a description of its intended use, and its level of autonomy in decision-making. Also, it may include details regarding the data (e.g., evaluation, training data), model performance, ethical considerations, and limitations.

Define and document the targeted population for model application [12], [66], [80], [82], [84]. In addition to defining the inclusion and exclusion criteria (both for the training data and for the targeted population), the team should ensure documentation for transparency purposes.

Consider how to communicate potential risks of an AI solution to end users and/or patients [12], [80], [81]. It is important to determine how risks will be evaluated, as well as how those risks should be communicated to end users and/or patients.

Assess factors pertaining to impact on patients [66]. Determine whether key performance indicators have been identified to measure the AI tool's impact on patient care (i.e., risks and benefits). Also ascertain whether a patient can opt out of having the AI solution used as part of their care, and whether patients will be informed about how their data will be used.

Consider regulatory/legal compliance issues related to the AI solution [85]. As with similar Safety considerations, this includes ascertaining whether regulatory and legal experts have been consulted and whether human factors and usability requirements have been evaluated. Additional questions include whether IRB approval and FDA submissions may be required for future application of the AI solution, and whether integration of the AI tool requires disclosures on the interface or consent from patients.

Consider the AI solution's impact on healthcare organizations [49]. Decide whether the AI solution will be reported within a registry, inventory, or centralized data platform, and identify whether the implementing organization has an established quality management system with which AI solution development must comply. In addition, determine whether independent quality reviewers and auditors have been identified and a method for reporting quality established. Finally, assess whether policies and standard operating procedures have been established, and if so, if they are transparent and accessible for relevant stakeholders.

Formulate a clear vision and define success measures [12], [80], [82]. Ensure that key performance indicators have been defined for the AI solution's intended use, and ensure that they will be tracked.

Establish specific goals, standards, terms, and conditions [49]. Determine if the AI solution's deployment goals can be quantified, and assess whether health and data standards (data provenance and representiveness) are defined. In addition, determine whether terms and conditions that comply with regulatory and ethical requirements have been developed, as well delineating any anticipated exceptions to those requirements. Finally, ascertain whether developer and implementer teams have a joint plan to align expectations with site-based requirements.

Consider stakeholder engagement [49]. Determine if key stakeholders have been identified and whether tracking of key stakeholder involvement has been established.

Security and Privacy: Stage 1

Ascertain whether the implementer organization has inventoried and documented its AI systems and solutions and mapped the data processed in connection with their use [3], [4]. Determine whether the organization has complete documentation of AI systems, including inventories of systems, solutions and attributes (e.g., documentation, links to source code, incident response plans, data dictionaries; AI actor contact information), as well as data maps for

data processing related to those systems. In addition, identify personnel responsible for documenting and maintaining AI system inventory details and determine whether cybersecurity and privacy risk assessments have been conducted on the AI systems.

Establish and maintain policies and procedures to manage AI privacy and security risks [3], [4]. Organizational stakeholders' policies and procedures should take into account mission priorities, risk tolerances, and legal and contractual obligations as well as the organization's external roles. Assess whether the developer team's organization understands and has documented the privacy and cybersecurity risks, as well as the legal and regulatory requirements, of its AI system in the context of the healthcare industry and its mission priorities and risk tolerances. In addition, ascertain whether there are risk management processes defined by privacy and cybersecurity policies in place for AI systems at the implementing team's organization.

Define the proposed use of AI systems in relation to specific mission/business objectives [3], [4]. This consideration aligns with similar considerations under the principle of Transparency. The implementer team should determine whether transparent processes and documentation exist to evaluate the purpose of the proposed AI solution (e.g., is there a specific task in mind? What is the funding source?) and how the solution helps the implementing team's organization meet its goals and objectives. Determine whether the use of the solution is consistent with the implementing organization's risk tolerance and whether the solution offers an appropriate way to achieve stated goals in light of identified risks.

Conduct initial privacy and security risk assessments on proposed AI systems [4]. Identify how known privacy and security risks are prioritized, and ascertain whether there are documentation and supporting rationales for risk responses, as well as a process for updating risk assessments throughout AI lifecycle stages.

Use Case-Dependent Considerations for Stage 1

Genomics AI Use Case (Precision Oncology with Genomic Markers)

This use case requires integrating clinical data, genomic insights, knowledge databases, and clinical trial findings to identify the best treatment for a patient. Precise documentation is crucial because datasets may favor specific groups, potentially impacting treatment outcomes. Transparency about these biases is essential, particularly considering how genetic differences can influence treatment effectiveness across various demographics. Additionally, it is vital to clarify the sources of knowledge databases and establish clear cut-off dates for clinical trial data inclusion. This ensures transparency and supports informed decision-making. Additionally, it is important to acknowledge that certain demographics may be underrepresented in clinical trials due to factors like income and race. Collecting this demographic data is critical for thorough analysis and to mitigate potential biases in treatment recommendations.

Predictive EHR AI Risk Use Case (Pediatric Asthma Exacerbation)

Given that the AI solution is integrated with and utilizes patient data from the EHR, it is necessary to ensure that privacy and security measures cover both the EHR and the AI application. The Asthma Exacerbation (AE) risk score application operates separately from the EHR but is accessible through it. This setup likely demands additional or specific privacy and security measures, such as requiring authentication into the EHR for accessing the application or additional authentication steps.

Generative AI Use Case (EHR Query and Extraction)

It is essential to clearly outline the problem and the necessity of the AI tool, which addresses the challenge of efficiently navigating vast and unstructured EHR data. This tool assists healthcare professionals in swiftly and accurately accessing patient information during clinical encounters. However, establishing specific goals and outcome measures for a general-purpose information extraction tool can be challenging. It is important to consider the specific compliance requirements of different jurisdictions. Healthcare systems often operate under diverse regulatory frameworks, highlighting the need to ensure that the AI system complies with local privacy laws and international standards such as GDPR or HIPAA. Additionally, plans should be in place for handling adverse events or AI system failures. This includes protocols for responding to situations where the AI system fails or provides incorrect information, along with strategies to mitigate any potential harm to patients.

These use cases are fully described in Appendix 1

Stage 2: Design the AI System

Usefulness, Usability, and Efficacy: Stage 2

Consider the usability of the AI solution [86]. This step includes determining whether the usability of the AI solution has been assessed and documented in the design phase, as well as whether human factors principles and usability heuristics have been explicitly considered and applied.

Implement methods to facilitate trust in the AI solution [63], [64], [65]. Because trust is essential for successful adoption and impact of an AI solution, it is important to document potential trust in the AI solution using a risk-benefit assessment. The team should ascertain whether the AI solution has undergone thorough robustness testing, and whether this process and its outcomes have been documented.

Assess how the tool will need to be tailored for the specific work context of the implementing organization [56]. This includes determining whether there is a description of the proposed development environment, and if an assessment evaluating differences between the development and implementation environments has been conducted.

Fairness: Stage 2

Consider how the choice of AI solution outcome(s) will affect bias and fairness [27], [87]. In order to determine whether a measure of real-world/clinical outcome (beyond AI model performance) has been defined and adequately justified, ask whether real-world/clinical outcome measures will be available for evaluation with sufficient time to assess the solution's impact and in a way that represents the target population. Also consider whether real-world/clinical outcomes will be compared for equality across all relevant socio-demographic subgroups.

Provide clear documentation of AI model development procedures, risks, and limitations related to fairness and bias [83], [88]. The team should identify any limitations to the interpretability and generalizability of AI system/outputs across the deployment setting(s) and in socio-demographic subgroups, including whether biases exist in AI model performance by subgroup or in retrospective data from different settings that are not addressable statistically or through procedural changes. If there are, consider how to document these limitations. In addition, consider whether there are unaddressable limitations in sample size, power for parity-based analyses, confounds, etc. If so, assess whether these have been clearly identified and documented as potential limitations/risks.

Consider appropriate and effective channels for end user feedback related to bias and fairness [89], [90]. This step includes assessing whether end users and/or implementers can provide timely feedback about performance, accuracy, or operational challenges of the AI

solution that have the potential to increase systematic harm or reduce benefit in ways that are biased or unfair. The team should also consider whether feedback strategies are simple, informative, and quick and easy to access and complete, as well as ensuring that feedback is reviewed in a timely manner, preventing existing issues from escalating or causing harm.

Consider stakeholder review and approval of the implementation process [38]. Key parts of this step include determining whether AI solution implementation procedures, risks, and limitations are reviewed and understood by all relevant stakeholders prior to moving to a pilot phase. The team should also identify any approval processes and evaluation criteria that would require updates/changes to the AI solution prior to pilot phase.

Safety and Reliability: Stage 2

Ensure that end users of the AI solution are able to control, direct and override recommendations as appropriate [39], [42], [91]. Key parts of this step include 1) determining whether the AI solution's end users will be able to control, direct, and intervene in basic AI system operations if there are safety concerns or important risks, and 2) identifying whether any override actions are recorded when an end user makes a decision that differs from the AI solution's finding or recommendation.

Ensure that a process is in place to manage ethical and legal challenges [36], [49]. This process should include asking the following questions: Is there a process for identifying errors that should be disclosed, along with a channel to disclose them, receive responses, and respond to the outcomes of disclosures? Have legal considerations been taken into account (e.g., Are ONC and HHS transparency and interoperability regulations observed where applicable? What happens if a model is not approved or if there is an FDA recall?)? Are there cases or lawsuits that end users and others who participate in the AI solution workflow should be aware of? Will patients be informed that AI is being used so that in case of adverse events, the health system is covered? Are there local laws for informed consent or safety reporting with implications for implementing the AI solution? Are mechanisms in place for disclosing safety issues, including to researchers? Is there information that should be disclosed to patients at other organizations where the AI solution is deployed, and is there a means to disseminate that information? Is an IRB submission necessary for research involving the solution's use on human participants?

Plan risk assessment methods from conception through to deployment of the health AI solution [5], [12], [14], [39], [49], [76], [77]. Risk management planning and assessment methods should be developed for the deployment of an AI solution using a risk-based approach to patient safety. Safety risks and potential for harm may involve errors or malfunctions related to AI system output, recommendations, supporting software/hardware, etc. Findings from requirements gathering, design, engineering and testing should be added to a risk management plan detailing potential adverse events and safety issues, as well as their causes, to the implementer and developer organizations, and how those issues will be corrected, also known as Corrective and Preventative Actions (CAPAs). Those actions should address risks and

opportunities for improvement, as well as prevention and reduction of harm, bias, and undesired results.

Appropriate implementation of clearly structured and consistently repeatable decision-making processes by implementer organizations can assure that efforts to minimize patient safety risk and promote patient safety have been considered. The implementer organization should have SOPs in place for risk management, ensuring consistency in decision-making for identified risks. Ensure that potential risks to safety or harm are captured for reporting to the developer and implementer organizations. Reports should include details about the rate of occurrence, apparent causes, whether those causes could be corrected, and any significant potential effects on patient care.

The risk management plan should be accompanied by the design of a patient safety focused process and a framework for the measurement, analysis and improvement of processes and the AI solution, including document control and records, configuration management and control, access controls, change management and managing outsourced processes. Document control and records management also serves to help communicate and preserve the rationale for why certain decisions related to the AI solution, e.g., related to patient safety or risk management, were made.

Assess whether the initial deployment of the AI solution constitutes human subjects research (HSR) [49]. In cases where a new AI solution has not been cleared by FDA, the team should identify whether deployment of the proposed AI solution constitutes human subjects research (HSR) via communication and consultation with the implementer team organization's IRB. If the solution deployment is determined to be HSR, ensure that IRB requirements are met.

Plan a monitoring process for adverse events (AEs) and serious adverse events (SAEs) [49]. When determining whether there is a monitoring plan for safety risks, including a breakdown by severity and frequency, the team should consider questions such as: Is there a common organizational standard for "adverse event" and "serious adverse event"? Are "serious adverse events" resulting from use of an AI solution tracked separately from "adverse events" and if so, are they tracked in a timely fashion? Are plans in place for sunsetting an AI system as needed, including conducting an associated safety investigation and triggering a back-up plan to enable associated health system workflows to continue functioning?

Ensure that AI models are labeled with transparent information about their development and limitations [42], [49], [91], [92], [93], [94]. Key parts of this step include determining whether potential limitations including dataset, model, and system constraints, as well as important details about accuracy, error rates, generalizability, and clinical implications are disclosed to the implementer organization. The team should also determine whether the implementer team will incorporate an explanation to the end user about why the AI solution made or suggested a particular decision.

In addition, the team should determine whether the developer team incorporates a method (e.g. Model Cards) to communicate to end users the fact that they are interacting with an AI. A related task is assessing whether the solution includes a plain-language explanation of how the underlying model was developed, its intended purpose, and its limitations and safety risks (e.g., type of model, description of dataset used to create the model, results from clinical studies, and which subpopulations were underrepresented in the training and test sets).

Finally, the team should evaluate whether transparency documentation includes a clear explanation of the model's limitations and clinical implications (e.g., error rates, contraindications, generalizability, reproducibility, robustness, etc.) determine whether a process exists for updating this documentation based on newly discovered limitations of local deployment in the implementer team's environment.

Establish clear reporting and recall procedures [95]. Reporting and recall procedures should identify and address flaws, biases, and safety concerns discovered during pilots in multiple health systems, ensuring timely notifications to developer organizations, relevant agencies, and all users of the AI system. Other important considerations include whether there is a defined threshold for reporting potential patient harm, and whether processes are in place to report on it. The team should also determine whether safety concerns meeting the defined risk threshold at one health system (thereby triggering a delay in the pilot phase and re-evaluation) are shared with the developer organization and regulator(s), as appropriate.

Enable clinical intervention and override by ensuring that the AI system is intelligible to end users [42]. This step entails determining whether the AI solution provides an explanation to end users regarding the specific rationale behind its decision(s).

When possible, design the AI system to keep a human in the loop to contest or override AI output, ensuring that a human-machine teaming model specifies how the user interfaces with outputs [42]. As indicated by the consideration above, it is important to clearly assess whether there is a human in the loop who can oversee, override, or contest the AI solution's output. If not, the team should determine what would be required to add a human to the workflow, or else implement additional quality control processes to assess AI system accuracy and safety.

Ensure that representative stakeholders (especially end users) are included in the AI model engineering and design process [12], [39]. Representative stakeholders, including end users, should be part of the AI system design process. This helps the developer team to identify potential safety risks to patients and develop corresponding mitigation plans. Relevant stakeholders include participants in the (clinical) workflow, who are key for identifying potential harm issues and risks related to including data elements, and also for clarifying data quality issues.

Transparency, Intelligibility, and Accountability: Stage 2

Justify the model design and architecture choice [66], [80], [81], [82], [84]. This involves assessing whether the AI system has been compared to other benchmarks. The team should determine whether all predictors used in developing/validating the model have been documented, including how and when they were measured and whether features adhere to meta-level requirements (e.g. principles). Developers should share information on methods and results of internal validation. The team should also identify whether the type of model and model-building procedures (including predictor selection) have been defined, while keeping IP in mind.

In addition, it is important to consider whether or why a simpler model is not sufficient or does not perform better than the model underlying the proposed solution. The team should also determine how model output for patients will be documented. For example: if the output is a score or percentage that can change over time, will previous scores or percentages be stored? Will patients have access to previously documented model outputs?

Consider decision thresholds [12], [81], [96]. Determine whether decision thresholds for the AI model have been established and whether they are clear and understandable to end users when they engage with the model output. A critical question is whether a mechanism exists for the development team or AI system to provide explanations to end users regarding the rationale behind specific decisions or recommendations made by the AI solution. When possible, the ideal situation would be for development teams to adjust thresholds based on local deployment environment. The team should seek to map AI outputs, and thresholds should be developed via a formal, normative procedure, employing a decision-theoretic analysis, where utilities or costs and benefits of action are considered under uncertainty.

Consider end user understanding of the model [12], [23], [81], [83]. For this step, the team should determine whether documentation of how to use the model exists and note whether it takes into account variability of end user knowledge and expertise. Further, they should ensure that users can reliably use the AI solution outputs to effectively and safely interact with the system. Also determine whether end users were involved in the model's development to ensure appropriate functionality and clinical fit at the implementer organization.

Consider issues of accessibility and explainability [83]. The team should determine whether the AI solution's performance has been assessed across all demographic groups. In addition, note the degree of explainability for the solution and its outputs and whether those can be measured. Consider whether transparency measures have been defined to provide different user-facing views of model outcomes (e.g., options vs. automatically ranked or triaged), thereby ensuring that interpretation bias is mitigated.

Security and Privacy: Stage 2

Evaluate privacy and security requirements for AI systems in the context of the implementing organization's understanding of risks [3], [4]. Privacy and security risks apply to individuals and operations as well as legal, regulatory, and contractual obligations. Determine whether there is traceability between AI system requirements and privacy and security risks and obligations. In addition, assess whether legal staff were interviewed to determine relevant security and privacy legal, regulatory, and contractual obligations and establish whether the risk management strategy includes evaluating risks both to individuals and to the implementer organization. Finally, ask whether privacy and security risk assessments during the problem definition and planning phase were reviewed for risks that can be mitigated by system requirements.

Protect development and production environments by securing user access [3], [4]. This step includes determining whether user access control policies and procedures (including for remotely connecting to the AI system's environment) establish a lifecycle of account management while incorporating the principles of least privilege and separation of duties. The team should also assess whether user access control records for the AI environment reflect account management according to those policies and procedures. Finally, evaluate whether the information flow configuration demonstrates that the AI solution's environment implements network protections such as segregation or segmentation.

Minimize privacy and cybersecurity risks through system architecture design and use of privacy-enhancing technologies (PETs) where applicable [3], [4]. A key part of this step is to ascertain whether the AI system's output identifies individuals or behavior, directly or indirectly. The team should assess whether the implementing organization has privacy attack mitigations such as differential privacy or other PETs in place in its AI environment. In addition, identify whether the AI solution's architecture mitigates privacy and cybersecurity risks, and whether audit log processes are in place to monitor data privacy outputs.

Establish mechanisms to incorporate contextual factors like privacy preferences into AI system design and implementation [3], [4]. Determine whether there are personnel responsible for incorporating contextual factors, such as individuals' demographics and privacy preferences, into AI solution design. Consider whether the implementer organization can describe the expected and acceptable context of use, including demographics, privacy interests or perceptions, data sensitivity and/or types, and visibility of data processing to individuals and third parties. Finally, the team should assess whether mechanisms used to incorporate contextual factors, such as surveys, focus groups, generative AI learning models, and user interactions, are adequate for the specific needs of the implementation.

Use Case-Dependent Considerations for Stage 2

Imaging Diagnostic AI Use Case (Mammography)

AI solutions involved in acquiring, processing, or analyzing medical imaging data are likely to be classified as software medical devices under the FDA's 2022 regulations [87]. Consequently, they will be required to obtain FDA clearance, meeting Criterion 1. Organizations implementing these solutions must obtain evidence from the developer demonstrating compliance with federal regulations. Alternatively, if organizations are involved in co-developing and deploying the model, they must ensure adherence to federal regulations throughout the process.

Claims-Based Outpatient AI Use Case (Care Management)

End users of the AI solution should have the capability to control, direct, and override recommendations as necessary. It is crucial to determine at which stage of the workflow human intervention should occur. For instance, end users might not have access to all levels of member data to assess model performance accurately. This is especially pertinent as risk stratification alone does not dictate final decisions, and additional assessment is needed for care management program enrollment. Assigning the responsibility for monitoring model drift (Stage 6) to the individual overseeing the need for human intervention may be more effective. Various end users with different qualifications, such as physicians, nurses, case managers, and behavioral health specialists, may utilize the model output for care coordination. Therefore, it is vital to ensure that instructions for use are tailored to each end user or are sufficiently comprehensive to facilitate consistent use across all user types.

Generative AI Use Case (EHR Query and Extraction)

It is important to justify the model design and architecture choice since different models (e.g., encoder-only like BERT, decoder-only like GPT, encoder-decoder like Flan-T5) impact the AI's understanding and response generation effectiveness. Assessing how well the system scales to manage large query volumes across multiple facilities is crucial, especially for expansive healthcare networks. Transparency regarding poor or missing data is paramount for these use cases. Users must understand that if certain information is not recorded in the clinical record, the AI will not be able to retrieve it. However, it is vital to recognize that the absence of data does not necessarily mean the patient does not have a particular condition or has not received a specific treatment. Protecting against unauthorized access and data leaks in the AI environment is imperative due to the sensitive nature of personal health information. Robust safeguards must be in place to prevent unauthorized access. Special attention should be given to the anonymization techniques used when training the models with sensitive patient data to mitigate any potential data leakage risks.

These use cases are fully described in the Appendix 1

Stage 3: Engineer the AI Solution

Usefulness, Usability, and Efficacy: Stage 3

Consider data quality and integrity [97]. Training and testing data source quality should be assessed to determine if it is sufficient for AI model and evaluation, including whether it is free from major errors or inconsistencies. This includes determining whether missing data types or data points have been appropriately addressed and aspects of the data that may lead to automation surprises for users have been addressed.

Consider the bias and fairness implications of the AI system, including during feature extraction [98], [99]. For this step, evaluate whether the data used for training and testing is diverse enough to allow assessment of its performance in patient subgroups. Also, determine whether they have been examined for potential biases related to factors such as age, gender, ethnicity, etc. A further consideration is whether measures are in place to ensure that the AI system's decisions are fair and unbiased.

Ensure that the availability of data used for AI model training matches deployment [100]. If an AI model is developed on retrospectively collected data, the team should determine whether all necessary inputs will be available at the time model output is generated (for example, diagnoses coded from notes are only available after a hospitalization has ended). The team should also assess whether all possible data sources for a given input have been accounted for (e.g., a cardiac ejection fraction measurement could be in a separate physician note, or different sites may differ in how they collect it, even if they use the same EHR software).

Fairness: Stage 3

Determine whether the use of protected characteristics or related features/proxies during AI model training and testing can be clinically justified [27], [101]. For this step, assess whether the AI solution and its underlying model explicitly or implicitly (due to highly correlated variables or proxies) use protected characteristics to make or recommend decisions, and if so, ask whether the process is clinically justified and necessary. If the use of protected characteristics, correlated variables, or proxies is clinically justified, the direction and magnitude of the effect of these features should be quantified. In addition, assess whether the contributions of protected characteristics to decisions improve fairness as predefined or improves the balanced allocation of resources, access, and outcomes in historically or currently underserved subgroups who typically experience poorer outcomes.

Assess for potential disparities between training and testing data and the target population [102]. It is important to consider significant representational disparities, like missing data, between the intended target population and the input or output distributions in the training or testing datasets. If there are significant disparities, they should be addressed.

Consider how relevant sociodemographic subgroups are defined and assess the availability of these data [102]. Determine whether training and testing datasets provide information on relevant socio-demographic subgroups, as well as representative data from the deployment setting, that will eventually allow for fairness evaluation between them.

Assess for potential bias/issues in data quality by relevant socio-demographic factors/subgroups or context [27], [103]. For this step, the team should evaluate whether differences in data quality are likely across deployment sites, especially for clinical data (e.g., MRI scanner type, type/method of heart rate measurements, type of assay used, etc.), and whether such differences on data distributions have been evaluated. Explore any statistical interactions between data quality or data type and relevant socio-demographic subgroups (e.g. are Black patients or older patients more likely to have different, missing, or lower-quality data, or have data on a 1.5 vs 3 Tesla MRI scanner, etc.?).

Evaluate the appropriateness of proxies and composite scores and their impact on fairness and bias [9], [27], [104]. If proxies or composite scores are used as inputs or outputs of the AI model, clarify whether they have been evaluated for bias across relevant socio-demographic subgroups. Consider whether their use (as inputs or outputs of the AI model) could result in unintentional exclusion or differential treatment of already disadvantaged groups (e.g., cost/utilization of as a proxy for deciding on advance care coordination). In addition, the team should determine whether there is a directly measurable quantity that could be used instead of a proxy or composite score. If not, consider whether training, testing, and deployment data have been evaluated for systematic differences in proxies or composite scores by relevant sociodemographic subgroups that could be related to issues with access, especially if the model is intended to provide care coordination, clinical care, or need-based services.

Examine the robustness of data representation [27], [105], [106], [107]. When considering overall robustness, it is important to ask if representative *and separable* data is available to training and test the model's handling of different scenarios and data variations. Additionally, the team should assess if cross-validation has been done using k-fold (with appropriate K defined given the sample size), as well as cross-validation with one subgroup left out.

Ensure that local data for model tuning is representative of the present population and setting [27], [108]. It is important to evaluate the model's performance and fairness based on data representing the population in which it will be implemented. It is also important to ask if the model has been tuned to a local population using retrospective or current data.

Consider the availability of information about the data used to train and test the model as well as its appropriateness for evaluating fairness and bias [27], [73]. If relevant socio-demographic subgroup/feature data are available in training or testing datasets they should be evaluated so that representativeness analyses can be conducted and reported by the developer team (if not already available/provided). Developer and/or implementer teams should also determine whether they can compare the representativeness of training/testing samples to that of

the intended population for AI solution deployment. In addition, if information on training and test data acquisition and data purpose are available from the developer team, these should be shared with the implementer team. Finally, the justification provided for data selection/curation should be evaluated.

Safety and Reliability: Stage 3

Ensure that AI model training data represents the deployment patient population [12], [49], [109]. As with Fairness, and important Safety consideration is data representation. Ensuring a high-quality model will require the team to assess whether the dataset is sufficiently large and representative of the patient population in which it will be used. The team should also assess whether population-representative data is available to avoid bias and safety issues when the model is deployed.

Ensure that data governance and appropriate systems are in place to monitor data and data quality [5], [12], [14], [35], [49], [109], [110]. Monitoring data quality and dataset drifts can help detect drift in model output and effectiveness and prevent downstream safety risks. For these reasons, it's important to determine whether there is a system in place to monitor data quality, latency, and security, as well as outcomes and drift. The team should also identify any data governance and change management plans to drive accountability and reduce safety risks, as well as roles and responsibilities that specify who will address issues as they arise, given that data input and AI model output deviations may create safety risks. Finally the team should consider whether thresholds for data quality have been established (i.e., the extent to which the AI solution will continue to be safe and operate if there is a defect noted in data quality).

Ensure that stakeholders can trace complaints, ethical concerns, and safety risks related to data quality [39], [42], [43]. Leadership and accountability in the implementer team's organization should be identified, as well as data governance, quality policies and internal audit processes. The team should also assess whether the AI system's data lineage and provenance are auditable by independent third parties and whether there is a data quality issue management plan for the AI solution, so that risks and issues can be continually identified, assessed and managed.

In addition, the team should clarify whether stakeholders can report complaints, ethical concerns, and safety risks due to data quality, and whether Corrective and Preventative Actions (CAPAs) for these complaints and other issues exist. The team should also clarify whether all involved in the development and deployment of the AI solution, including third-party vendors and consultants, understand their roles and responsibilities as indicated in the change management plan pertaining to data governance, data engineering and data quality.

Apply clear inclusion/exclusion criteria for the targeted patient population [49]. AI models should have clear inclusion / exclusion criteria. If populations lack detailed exclusion criteria, additional detail may be needed, as this may limit validity. In addition, the team should identify

whether there is a protocol for automating or altering the inclusion/exclusion of deployment populations for whom the AI solution does/does not apply.

Implement proper access controls and audit trail mechanisms for stakeholders who will use the data and the health AI solution [39]. In overlap with Security and Privacy considerations, it is important for Safety's sake to maintain user access control records for the AI environment, showing that the system is managed according to established policies and procedures. This entails an audit trail and governance structure that can ensure compliance with regulations, detect breaches, and allow for independent review of who can access the AI solution.

Transparency, Intelligibility, and Accountability: Stage 3

Consider data security and scalability planning [82]. It is important to be transparent about data privacy and security requirements as defined by both the developer and implementer teams. This also includes assessing whether the AI model is engineered to be monitored for data breaches or violations, and whether system performance can be measured as complexity increases.

Ensure transparency in data monitoring [14], [82], [84]. Determine whether there is a committee or group that monitors training, testing and deployment data and whether their roles and reporting structure are established and documented. If such a committee/group has been deemed unnecessary, the team should find out if the justification for that decision has been documented.

Include socio-demographic information with diversity details, ensuring transparency for the sake of the target population [80], [81]. Determine whether the AI model training data is representative of the intended deployment population, and whether the correlation between them is documented. In addition, consider whether comorbidities and sociocultural influences have been accounted for.

Document data provenance, and specify the limitations of the data [12], [66], [80], [81], [82]. It is important to keep a record of training and testing data, including information on its origins, transformations, usage, and dependencies exist. Along with that, it is important to document data limitations (e.g. incompleteness, noise and errors, temporal bias, sample size, etc.).

Ensure documentation of data lineage [84]. For this step, the team should identify whether the data pipeline has been documented in a way that allows decisions about the AI model to be traced back to specific points and transformations. The team should also determine whether there is a plan for regular audits of data lineage to ensure that it remains accurate and current.

Implement version control for datasets [81]. Determine whether there is a tracking process to maintain dataset version control, as well as a process for notifying end users of changes made after deployment.

Consider the impact on patients and the potential need for consent [111]. Determine whether the extent to which patients have access to information about the AI system and its output has been justified and documented.

Ensure transparency into rationale for manipulating data [49]. For this step, assess whether any types of data manipulation used (e.g., feature engineering, data cleaning, text preprocessing, etc.) have been justified and documented.

Security and Privacy: Stage 3

Implement controls for privacy and security requirements for the AI system [3], [4]. Privacy and security requirements are important in order to enable the implementer team's organization to address privacy and security risks to individuals and operations as well as legal, regulatory, and contractual obligations, the team should determine whether there is traceability between AI system requirements and privacy and security risks and obligations. In addition, find out whether legal staff have determined whether implementer team requirements meet relevant security and privacy legal, regulatory, and contractual obligations. The team should also assess whether the risk management strategy includes evaluating risks to both individuals and the organization, and consider whether privacy and security risk assessments will be conducted again to assess whether the implementation has altered risks.

Ensure that data management policies and capabilities are informed by privacy and cybersecurity risks, AI system requirements, and individuals' privacy preferences [3], [4]. This step includes determining whether policies address the management of data processing authorization and revocation, including individual consent where appropriate. Also note whether policies address how data will be managed to minimize privacy and cybersecurity risks and meet AI system requirements, including data retention and data quality management.

In addition, consider whether policies address the management of individuals' privacy and data processing preferences, and whether the AI system enables compliance with those policies. Also assess whether there are means for obtaining feedback about privacy preferences (e.g. surveys, focus groups) and if so, what the results are. Finally, note whether stakeholder privacy preferences were included in AI system design objectives.

Ensure that protections are implemented against unauthorized access and data leaks in the AI environment [3], [4]. Determine whether AI system data stores protect confidentiality and data integrity, as well as whether the AI system network protects the confidentiality and integrity

of data transfer. Also note whether AI system user access and network controls maintain separation between AI development and testing environments.

Evaluate whether data inputs and provenance enable accuracy, ensure completeness, and facilitate the management of bias in datasets [3], [4]. As part of this step, the team should identify documentation regarding how, from what source(s), and under what circumstances data elements were acquired for the AI solution (including manner and mechanism of consent where appropriate). Also identify those persons involved in the data collection, and the categories of individuals whose data are being used. Note whether there is anything about the composition of the dataset or the way it was collected or processed that might affect future uses, and whether there are tasks for which the data should not be used. Finally, consider whether the dataset should be updated and if so, after what interval(s)?

Protect development and production environments by securing user access [3], [4]. Assess whether user access control policies and procedures for remotely connecting to the AI system environment establish a lifecycle of account management while incorporating principles of least privilege and separation of duties. Also note whether user access control records for the AI system environment confirm that account management is consistent with user access control policies and procedures. In addition, determine whether the information flow configuration for the AI system environment implements network protections such as segregation or segmentation.

Use Case-Dependent Considerations for Stage 3

Imaging Diagnostic AI Use Case (Mammography)

During testing, it is crucial to assess barriers to AI tool utilization in the local environment. This includes evaluating how the AI is accessed and whether the interface is user-friendly. For instance, excessive clicks or the need to manually launch different programs to view AI output can impede user adoption in the local setting. Addressing biased training data is especially critical for imaging tools. Some implementers may require fine-tuning of established products because their population is predominantly non-white, which can affect the effectiveness of imaging tools like mammography.

Generative AI Use Case (EHR Query and Extraction)

Document control and records management are essential for preserving the rationale behind decisions made during the engineering of the solution, particularly concerning patient safety and risk management related to the AI tool. Implementing a solution that audits the tool's use by clinicians, tracking EHR document access and the questions and answers provided by the system, is critical. This access could influence clinical decision-making and may be subject to litigation or adverse event alerts.

Claims-Based Outpatient AI Use Case (Care Management)

End-users, such as care coordinators, case managers, and nurses, rely on the model's output, specifically risk levels, for their tasks. They may not possess knowledge about the data used in the models, as it is unnecessary for the proper use of the output. It is crucial to define the term "end-user" and specify their role in the workflow. Certain subgroups may experience more missing data due to various factors, such as data sourcing or limited access to digital records or health services. This can lead to incorrect categorization as low-risk or result in individuals not receiving a risk classification. It is essential to address these cases equitably by identifying which subgroups are more susceptible to data gaps, finding ways to address these gaps, and managing care coordination efforts for individuals lacking key data necessary for appropriate risk categorization.

These use cases are fully described in the Appendix 1

Stage 4: Assess

Usefulness, Usability, and Efficacy: Stage 4

Assess how the AI solution will integrate into workflow [59], [60], [61]. Once again, it is important to ensure that a workflow integration assessment has been completed and documented. It is also important to assess how the AI solution accounts for the flow of people and tasks through physical and digital environments and how the solution may hinder patient-clinician interactions. This includes determining whether relevant team activities (e.g., clinician-clinician, patient-clinician interactions) have been studied and the possible effects of the solution on team activities have been assessed. In addition, determine whether end users and others whose work will be affected by use of the AI solution have been defined or identified.

Reassess whether the problem defined in stage 1 is addressed by the AI solution [12], [55], [56], [57]. This includes reevaluating whether the AI solution addresses the stated use case, is consistent with organizational objectives, and whether it potentially improves the standard of care or existing practice.

Reevaluate the usability of the AI solution [12], [86]. This step includes determining whether the usability of the AI solution has been assessed and documented in prior stages, as well as whether human factors principles and usability heuristics have been explicitly considered and applied.

Ensure that there are methods to facilitate trust in the AI solution [23], [63], [64], [65]. Because trust is essential for successful adoption and impact of an AI solution, it is important to document potential trust in the AI solution using a risk-benefit assessment. The team should ascertain whether the AI solution has undergone thorough robustness testing, and whether this process and its outcomes have been documented.

Assess how the tool will need to be tailored for the specific work context of the implementing organization [56]. This includes an assessment evaluating differences between the development and implementation environments.

Fairness: Stage 4

Consider model evaluation and calibration as it relates to fairness and bias across relevant socio-demographic subgroups, locations, or contexts [112], [113]. This step includes assessing whether counterfactual tests are conducted both with and without relevant sociodemographic subgroups in order to evaluate AI model performance. In the case of predictive AI models, the team should determine whether model calibration has been evaluated and documented across the whole test set and across sites/settings/subgroups.

Consider sample independence and fairness [114]. Assess whether training and test datasets are independent, and note whether the AI system is calibrated by producing outcomes independent of protected classes such as race, gender (or their proxies), disability, or other variables that correlate highly with protected classes.

Consider model performance, parity, and balanced allocation, access, or outcomes across relevant sociodemographic subgroups [27], [115]. Determine whether measures of parity, overall accuracy, and fairness are chosen, by accounting for the scope, degree, and direction of impact that errors or inaccurate predictions can have on individuals or subgroups. Determine whether measures of fairness are consistent with any definition of fairness stated during the problem definition and planning phase.

Consider measures of model performance and impact, particularly regarding fairness, bias, and balanced allocation of resources, access, and/or outcomes (as appropriate) beyond accuracy, sensitivity, and specificity [38]. Note whether a plan is in place and data available for evaluating how the AI system may improve the balance of resource distribution, access to care, clinical operations, and/or real-world clinical outcomes.

Safety and Reliability: Stage 4

Evaluate performance and safety at the local level, even if impact and safety have been evaluated for a given AI solution in other populations [49]. In order to inform risk management practices for the deployment of the AI solution, human factors should be integrated into safety and harm assessments. Even in instances where impact and safety has been demonstrated for a given AI solution on another population or in a different setting (e.g., for FDA-cleared AI technologies), that ensure performance and safety for the implementer team's site has been shown. Assess whether the AI solution has been deployed on a "test population" or in a test environment in the setting where it will be deployed and evaluated for safety and efficacy on the local target population.

Ensure that risk management and assessment methods are in place [5], [12], [14], [39], [49], [76], [77], [110]. Risk management planning, assessment methods, and risk mitigation strategies should be developed for the deployment and use of an AI solution using a risk-based approach to patient safety, as described in Stage 2. In the risk management plan, potential risks should be captured and enumerated at all prior stages of the lifecycle, as early as Stage 1, engaging all stakeholders involved in the AI solution development, deployment and use to have a holistic view of how the AI system may cause harm to patients.

During the use of the AI solution, any risks to safety or observed risks of harm should be triaged and reported to the implementer team and, in turn, to the developer team or developer organization, when applicable. The risk management plan should account for the risk measurement, analysis and improvement of AI-related processes, and it should include CAPAs addressing risks and opportunities for improvement, as well as prevention and reduction of harm,

bias, and undesired results. This should be accompanied by the design of a patient safety-focused process with clear thresholds for what should be reported. A feedback loop of end user disagreement with the output can ensure consistent detection of issues and defects for continuous improvement.

Additionally, a process should be in place to detect patterns of patient harm associated with a given AI model and report harms to the developer team, and a process should be in place for determining if the Al should continue to operate, needs to be refined, or should be sunsetted in the event of safety issues and/or poor outcomes.

Target verification and validation (V&V) activities toward the safety-critical nature of the AI system [12], [39], [92], [116]. Verification and validation activities should encompass both the clinical user and the environment in which the AI system is implemented, including factors such as whether the solution is properly installed, instructions for use are correct, and software safety elements work properly (i.e., user acceptance testing [UAT]). Basic software development lifecycle best practices such as validation are needed to ensure end-user acceptance for implementation [109]. Acceptable failure behaviors ('failsafes') in the clinical environment should also be established, taking into account workflow, environment, and stakeholder considerations.

In addition, AI systems should be judged on their implementation/clinical results, noting that users' level of trust and safety considerations influence clinical performance of the system. To that end, where applicable, determine whether structured human factors testing on a subset of patients has been conducted. Assess whether UAT and clinical evaluation (chart reviews, etc.) have been performed to demonstrate that clinicians support the implementation of the AI system, where applicable.

Ensure quality and transparency of validation methods and results [49], [92], [110]. Determine whether rigorous evaluation methods are used and whether explanations are provided to end users regarding validation methods and subsequent results (e.g., training population data, model performance based on sociodemographics, etc.). In addition, ascertain whether there is a protocol for disclosure of errors or hallucinations, accompanied by an explanation of implications for users.

Transparency, Intelligibility, and Accountability: Stage 4

Report on AI solution effectiveness to end users and key stakeholders [12]. Determine whether a preliminary study of AI solution effectiveness has been conducted and reported. This may include an evaluation of end users' and key stakeholders' understanding of actions based on AI model output can be measured, in order to ensure consistency with AI solution's intended use and its identified limitations.

Establish specific goals, standards, terms and conditions [49]. This step includes determining whether performance goals for the AI solution can be quantified and whether health and data standards (data provenance and representativeness) are defined. Determine if terms and conditions are in place that comply with regulatory stipulations. In addition, the implementer team should identify whether consent to use patient data has been obtained. Finally, it is important to ensure that a joint plan has been implemented between developer and implementer teams to align expectations with site-based requirements.

Define roles and foster trust and transparency [23], [49]. Evaluate whether assigned roles and responsibilities foster transparency and trust in the AI system and whether adherence to roles and understanding of the system among users and stakeholders can be measured. In addition, note whether there are documented justifications for AI model logic available for end users to communicate to patients (when applicable).

Consider data security and scalability planning [82]. Establish whether data privacy requirements, security requirements, and a plan for scalability have been defined. In addition, determine whether the AI solution is being monitored for data breaches or violations, and also whether performance can be measured as its complexity increases.

Consider accessibility, human-machine teaming and explainability [23], [83]. Determine whether the AI solution's service performance has been assessed and reported on for parity, and whether the results are acceptable according to external and internal standards. In addition, the team should assess whether the solution's outputs are usable and explainable, and whether transparency criteria have been defined for different user-facing views of model outcomes (e.g., providing options versus automatically ranking or triaging). Finally, determine whether the solution adheres to accessibility standards.

Consider the downstream impacts of the AI solution [12]. Assess whether the AI workflow and potential downstream impact and risk of the AI solution's implementation and use has been evaluated.

Consider risk, change, and competitive analysis [12], [81]. Determine whether a risk analysis and risk mitigation strategy has been established, as well as whether success in risk mitigation can be appropriately monitored and reported. Also note whether potential risks have been identified and whether AI system adaptability can be measured. Finally, determine whether a competitive analysis has been performed and if risk mitigation success can be compared against competitors.

Incorporate user feedback and documentation [12]. It is important to have a process for collecting and documenting user feedback. It is also valuable to have a process for measuring user satisfaction – both with the AI solution and with the clarity of the information provided to them.

Consider how to report on performance metrics, confidence intervals, uncertainty, and fairness and bias audits [12], [27], [80], [110]. Note whether performance metrics for the AI model have documented and explained. As part of this step, also determine whether confidence intervals are reported, including uncertainty when possible. Evaluate whether AI model threshold selection has been justified and note whether a risk assessment plan has been developed, documented and conducted.

Consider contingencies pertaining to testing data and generalization [66], [81], [82]. Determine whether the AI model has been tested on out-of-distribution samples. Determine whether the AI model has been tested on a sample size corresponding to its target for general deployment.

Security and Privacy: Stage 4

Ensure that the workforce is appropriately trained regarding their cybersecurity and privacy roles and responsibilities [3], [4]. For this step, the team should determine whether there is a training curriculum and related materials for the AI solution, and if so, whether it provides for training for specialized roles. In addition, note whether there is a process for updating training (and if so, at what frequency?) as well as procedures for documenting completion of training.

Assess implemented controls to determine whether they are performing as intended [3], [4]. Before entering the pilot stage, it is important to ensure that there is traceability between AI system requirements and privacy and security risks and obligations. It is also important to ensure that legal staff have been consulted so that the organization meets all legal, regulatory, and contractual obligations. Privacy and security risk assessments should be reviewed for risks that can be met by system requirements.

Identify third-party service providers and their roles in the AI environment and ensure that they are bound to privacy and cybersecurity controls in contracts and subject to periodic audits [3], [4]. For this step, the team should first ascertain whether a risk assessment has been performed on third-party solution providers in the AI system environment. The team should also determine whether the implementer team's organization has policies and procedures that require third-party suppliers to meet specific privacy and cybersecurity objectives.

In addition, the team should identify any records of scheduled third-party audits or audits performed on third parties in the AI system environment. If a third party created the system or some of its components, determine whether sufficient documentation at an appropriate level of explainability or interpretability exists. Identify any personnel responsible for assessing sufficiency of third-party systems or components, and determine whether processes are in place for third parties to report potential vulnerabilities, risks, or biases in the AI system. Finally, note

whether there are processes for mitigating concerns raised by third-party AI systems or components.

Use Case-Dependent Considerations for Stage 4

Clinical Ops & Administration AI Use Case (Prior Authorization with Medical Coding)

Evaluate protections against unauthorized access and data leaks in the AI environment, given the sensitivity of personal health information. Prior authorization relies on system interoperability, posing privacy and security challenges due to increased communication between multiple systems and potential access to sensitive data. Automation and AI integration aim to align the prior authorization process with clinical workflows by standardizing communication between EHRs and payor systems. User education is crucial for understanding automated functions, as current manual processes require extensive training. Human oversight is vital for ensuring fairness in the prior authorization process. Transparency regarding training data, algorithm parameters, and the approval and denial process is essential for maintaining fairness. Human involvement should be validated during Stage 4 to ensure a balanced approach.

Claims-Based Outpatient AI Use Case (Care Management)

In addition to training on risks or limitations, it is crucial to educate individuals on actionable steps they can take when faced with specific limitations or risks. People often struggle to act on information without clear associated actions and may ignore it as a result [117], [118]. End-user testing is particularly valuable for refining risk classification models, as providers and care managers have comprehensive knowledge of patients over time, which may not always be fully captured in the input data for models. There is evidence suggesting that care coordinators or providers may identify individuals as high risk even if the algorithm does not, due to various factors. Over-reliance solely on the model output for eligibility by payors may result in patients with complex care needs being overlooked. Therefore, it is essential to consider additional factors beyond the model's output when determining eligibility [119].

Generative AI Use Case (EHR Query and Extraction)

Large Language Models are complex technologies that may be challenging for users and patients to understand fully. Ensuring patients grasp the tool's limited scope, which is solely for augmented information retrieval and not clinical decision-making, presents a difficulty. Additionally, due to the natural language interface, it is important to educate clinicians on how to use the tool appropriately, despite the system including a module to prevent misuse. Safety concerns often revolve around two main issues. The first is automation bias, where clinicians may rely too heavily on the tool's output without verifying the supporting evidence from the actual clinical record. This blind acceptance of retrieved information could lead to safety issues, especially if false positives (hallucinations) are present. This concern should be addressed in the considerations. The second is false negatives, where the greatest safety risk arises when critical clinical information, such as severe allergies or contraindications, present in the clinical record is not retrieved by the system. In such cases, clinicians may assume the patient does not have the issue, potentially leading to adverse outcomes. It is suggested to include this consideration for AI systems with broad potential scopes like this one.

These use cases are fully described in the Appendix 1

Stage 5: Pilot

Usefulness, Usability, and Clinical Efficacy: Stage 5

Ensure that end users receive clear communication about AI solution capabilities to establish and maintain trust [23], [63], [64], [65]. Assess whether users understand the capabilities and limitations of the AI solution. In addition, determine whether the potential for trust in the AI solution has been quantified and documented using a risk-benefit assessment.

Assess whether the anticipated benefits, risks, and costs of the AI solution match the actual benefits, risks, and costs when used in the clinical environment [12], [62]. This step involves evaluating whether error rates and response rates of the underlying workflow improve after the AI solution is implemented. Also assess whether the AI solution is superior to standard of care and note whether relative benefit is documented.

Re-assess the usability and effectiveness of the AI solution [25], [86]. Take note of whether the usability or effectiveness of the AI solution changes when deployed in an actual clinical environment (e.g., user efficiency, effectiveness, satisfaction).

Ensure that there is a process for if/when a clinician disagrees with the AI solution's output [120]. Clarify whether there is a plan in place to manage clinician disagreements with the solution's output (e.g., human in the loop, human override, etc.).

Assess the actions users take after interacting with the AI solution [12], [121], [122]. For this step, note whether tasks involving the use of the AI solution are adequately supported, and if so, whether these tasks are different than anticipated.

Fairness: Stage 5

Assess how choice of real-world/clinical outcomes impacts bias and fairness [87]. Determine whether real-world/clinical outcomes have been quantified beyond AI model performance and consider whether such outcome measures are available for evaluation with adequate time and in a way that represents the target population. Also assess whether real-world/clinical outcomes are compared for equality across all relevant sociodemographic subgroups.

Consider the representativeness of the pilot site and approach/method and its impact on bias and fairness [123]. Ascertain whether the pilot population, site, department, or program adequately represents the entire population in which the AI solution will eventually be used. Assess whether the method/definition of the pilot could disproportionately include or exclude a sociodemographic subgroup, as well as whether the impact of such inclusion or exclusion has been evaluated.

Consider how humans will interact with the AI solution and whether operational processes or workflows may introduce unintended bias [27], [124]. At this step, determine whether data are available and methods defined to evaluate whether the AI solution is being used as intended by end users and whether variability in end-user behavior impacts treatment or outcomes of specific socio-demographic subgroups. Also evaluate whether the integration of the AI system into the workflow has been proactively designed to maximize intended use, develop trust, offer ease of use, and minimize inappropriate (and potentially harmful) automaticity in decision-making. The team should also examine tendencies in user decision-making to see if they introduce differential outcomes for sociodemographic subgroups (e.g. because of automaticity, lack of trust in AI solution, etc.).

Safety and Reliability: Stage 5

Using a risk-based approach to patient safety, implement risk management, assessment and mitigation methods during the deployment of the health AI solution [39]. Risks should be identified and documented throughout the pilot deployment of the health AI solution, risk assessment methods should be in place, and a mechanism should exist to act upon them. Actions should address risks and opportunities for improvement, and the prevention and reduction of harm, bias and undesired results. Document control and records management also serves to help communicate and preserve the rationale for why certain decisions related to the health AI solution, e.g., related to patient safety or risk management, were made. A feedback loop can ensure consistent detection of issues and defects as well as disagreement with the AI output, and a triage process can facilitate appropriate reporting, continuous improvement and monitoring. Moreover, risk management plans should be frequently updated to reflect safety risks and issues, their causes, CAPAs, and mitigation strategies from the pilot phase.

Additionally, trends and patterns of patient harm associated with a given AI solution inform the frequency of those risks within the risk management plan and harms are reported to the developer team. A process should be in place for determining if the Al solution will continue to operate, needs to be refined, or should be sunsetted in the event of safety issues and/or poor outcomes. Risks to patient safety or observed instances of harm (including indirect harm) are reported at a predetermined frequency to the developer and implementer organizations.

Document control and records, configuration management and control, access controls, change management and managing outsourced processes are maintained. There should be processes to manage risk arising from changes to the system, environment, and data. Attention to detail is critical in areas underlying the implementation of the algorithm – a simple data overwrite can potentially lead to an adverse patient safety impact.

Maintain a monitoring process for adverse events (AEs) and serious adverse events (SAEs) [14], [49]. Serious adverse events (SAEs) resulting from AI system use should be tracked, separately from adverse events (AEs), and in a timely fashion. Defining a common standard for SAEs like CDISC for clinical trials and coming up with a common risk framework for health AI

(like SaMD risk stratification from IMDRF and patient safety reporting) would be valuable. In lieu of that, a common standard definition at the health AI solution level SAEs and at the implementer team's organization level is key, as the latter are where the clinical risk is defined within the workflow.

Determine whether there is a monitoring plan for safety risks, as well as a breakdown by severity and frequency. From a reporting perspective, ascertain if there is a common organizational standard for AE and SAE. Assess whether SAEs are being tracked separately from AEs in a timely fashion. Determine if there are plans in place for sunsetting the AI system if needed (and triggering a back-up plan) and initiating a safety investigation.

Implement a clearly structured, transparent, and consistently repeatable decision-making process to minimize risk, thus providing confidence that patient safety has been considered [5], [12], [39], [42], [43], [49], [76], [77], [91]. This emphasizes effective communication and mutual understanding between the health system and developer organization regarding potential risks. From a collaborative vantage point, it is vital to manage risks arising from changes to the AI system, environment, and data.

Implement measures to mitigate automation bias [42], [49], [92], [125]. Identify whether the potential for automation bias is described in the risk management plan, training materials, and/or interface use instructions, along with mitigation strategies. Assess whether it is possible to measure automation bias, and if such measurement is included in the risk assessment process (for example, determining whether the incorrect AI system output can be detected and how it may have potential impact on subsequent decision making).

Establish robust reporting and recall procedures for promptly notifying developer organizations, relevant agencies, and all users when safety concerns are discovered [14], [95], [126]. During pilots and implementation by multiple organizations, safety concerns and issues (bias, etc.) may be discovered. The implementer team should clarify how these issues are reported to the developer team so that the responsible organization, the FDA/other relevant agencies, and all customers using the AI solution so notified in a timely manner. The team should also determine how recalls/corrections will happen and assess whether there is a defined threshold for reporting potential patient harm and processes in place to report on it. If safety concerns meet the defined risk threshold at one pilot site, triggering a delay and re-evaluation, determine whether this information will be shared with the developer organization and regulators. Finally, developer teams should identify whether a process exists for sharing recalls and corrections with implementer teams.

Ensure that there is a continued human factors evaluation at the early stage of clinical implementation [12]. This step includes determining whether user acceptance testing is both qualitative and quantitative, as well as whether human factors are captured as part of the assessment of harm during the pilot deployment of the AI solution. In addition, consider whether

end users and other stakeholders in the workflow, such as patients, are included in the pilot human factors assessment.

Establish a process to regularly review the AI solution's relevance and potential obsolescence during its deployment [49], [77]. Because it is important to review whether the solution becomes "outdated" or no longer medically relevant at frequent intervals throughout its deployment, the team should determine whether there are processes established to investigate the clinical relevance of the solution (or input variables) and ascertain whether better AI methodologies may be available. In addition, note whether there is a process to identify incorrect "knowledge" and to determine when a newer AI system should be used.

Transparency, Intelligibility, and Accountability: Stage 5

Consider the system's capacity to handle errors and increasing data volume [12]. Evaluate whether the robustness of the AI solution's error handling, mitigation strategies, and resilience to increasing data volume can be monitored.

Ensure education/training for end users during small-scale implementation [12]. Determine whether educational resources and/or training courses have been developed and distributed to inform end users about functionality and limitations of the AI solution. In addition, note whether a particular "point-person" will champion training, implementation and follow-up.

Identify a method for ongoing audit monitoring, ensuring that accountable parties are aware and capable of taking any required mitigation steps [12], [14], [81]. For this step, assess whether key stakeholders/end users are aware of required mitigation steps. Note whether a process has been established so that key stakeholders/end users can identify and report any unforeseen, unintended, negative, and/or adverse outcomes and whether a process exists for determining whether such outcomes have been sufficiently assessed during the pilot. In addition, clarify whether there is a method for contrasting findings/recommendations of the AI solution and those of the end user. Finally, determine whether accountability for decision-making has been defined and legally vetted and a post-deployment monitoring strategy established.

Consider end user experience [49]. Determine whether user experience and interactions with AI solution output have been assessed (through focus groups, surveys, follow up studies, research, etc.) and note whether there is a process to collect, assess and implement user feedback.

Consider continuous reporting methods [84], [110]. Determine whether continuous monitoring and reporting processes are in place for subpopulations, especially vulnerable ones. Clarify whether it is possible to detect and document AI model drift to ensure that any potential safety, efficacy, and ethical issues are addressed through a response plan.

Ensure that model limitations are communicated to end users and patients [80], [81]. Determine whether end users have been educated/trained on the AI system's intended use. In addition, clarify whether the AI solution's limitations have been documented and whether that information is accessible to end users and/or patients.

Ensure transparency into existing clinical trials [12], [66], [80], [82]. Determine whether reporting guidelines have been followed to report clinical trial results.

Security and Privacy: Stage 5

Clarify whether stakeholder privacy preferences are included in algorithmic design objectives, and if outputs are evaluated against these preferences [4], [5]. Determine whether the implementer organization has performed a privacy risk assessment (or similar exercise) on AI systems to understand stakeholder privacy preferences. Also assess whether the AI system is analyzed to help align with stakeholder privacy preferences either formally (i.e., audits, Data Protection Impact Assessment) or informally (i.e., committee meetings).

Ensure that audit log records are determined, documented, implemented, and reviewed in accordance with policy, incorporating the principle of data minimization [4], [5]. For this step, determine whether the frequency of AI system input/output and processes regarding user access audit logs have been established and implemented. In addition, note whether the implementer organization reviews AI system audit log records, and clarify whether audit log processes capture only the minimum of data needed.

Verify that configuration change control processes are established and in place [4], [5]. Determine whether the organization formally records and stores changes to the AI solution's deployment environment. Clarify whether configuration change records, including details of the change, can be traced to the owner. Assess whether the implementer organization and third parties can easily share configuration changes with each other.

Ensure that an incident response plan is in place [4], [5]. Determine if the implementer organization has developed and documented an incident response plan that is updated with appropriate frequency.

Verify that delivery and resilience requirements of critical AI services are understood and established [4], [5]. For this step, determine whether the implementer organization has documented a contingency plan for critical AI services, and if so, whether the organization has trained personnel on contingency plan implementation responsibilities. In addition, ascertain whether the organization's records document that the contingency plan has been tested.

Note whether privacy and cybersecurity risk for the AI solution is examined and documented, and whether mechanisms are in place to mitigate risks during deployment [3], [4]. For this step, note whether relevant staff have been interviewed to determine security and privacy risks and clarify whether the risk management strategy includes evaluating risks to both individuals and to the organization. In addition, note whether privacy and security risk assessments have been reviewed for risks that can be mitigated by AI system requirements and determine whether data management, security, and privacy controls are in place as part of organizational data governance policies. Finally, evaluate whether the organization has established corrective actions to enhance the quality, accuracy, reliability, and representativeness of the data.

Determine whether mechanisms have been established to incorporate contextual factors, including individuals' demographics and privacy preferences, into AI system design and implementation [3], [4]. Identify whether there are personnel responsible for incorporating contextual factors into AI system design. Also note whether the implementer organization has determined the expected and acceptable context of AI system use, including demographics and privacy interests/perceptions, data sensitivity and/or types, and visibility of data processing to individuals and third parties. Assess whether mechanisms used to incorporate contextual factors are performing as intended. These mechanisms could include surveys, focus groups, generative AI learning models and interactions with users, etc.

Use Case-Specific Considerations for Stage 5

Predictive EHR AI Risk Use Case (Pediatric Asthma Exacerbation)

After integrating end user preferences and feedback into the design and development of the AI solution, it is vital to assess its performance in practice. When presenting prediction outputs as high or low rather than numerical values or percentages for pediatric asthma care, supporting explanations should accompany them to aid clinician decision-making. However, this high/low stratification may lead to automation bias, misinterpretation, or misuse. If updates are required for the Asthma Exacerbation (AE) risk score model or associated data-processing algorithm, thorough testing and validation are necessary. Communication with clinicians and updating documentation, such as model cards, will likely be needed. It is essential to be transparent about the intended population and purpose of risk models, such as the AE risk model, for end users and patients. Given that the AE risk score is intended to support clinical decisions rather than serve as a diagnostic tool, clinicians must understand why the model made a decision to inform treatment decisions. Transparency regarding the reasons behind high or low classifications is crucial for clinicians to make informed treatment decisions. Users need context about the meaning of high/low AE risk scores and the contributing features to independently interpret and assess the model output.

Generative AI Use Case (EHR Query and Extraction)

The tool is intended solely for information retrieval purposes. There is a risk that patients may misunderstand the system's capabilities and expect it to make decisions or influence clinician judgment beyond improving information access. The varying levels of risk associated with the tool's use should be considered based on specific queries or intended use, necessitating end-user testing and documentation, which may burden clinicians. Mitigating hallucination (confabulation) is relatively straightforward and is addressed by the system, as it always provides the actual content supporting the AI response. The greatest risk lies in false negatives, where certain facts in the clinical record are not identified. Instructions to clinicians should account for this risk, advising them to use additional safety measures to confirm the absence of critical data, such as a severe allergy to an antibiotic.

Genomics AI Use Case (Precision Oncology with Genomic Markers)

Genetic datasets often favor specific ethnicities and sample types, highlighting the need for transparency regarding the represented subgroups in the input data. However, AI systems can aid in achieving demographic parity. As certain subgroups are underrepresented in precision genomics cancer trials, AI systems can measure this data and provide insights to enhance trial participation. Nevertheless, accessibility remains a challenge, particularly for resource-constrained sites lacking personnel with expertise in bioinformatics, molecular pathology, and genetics. An essential consideration is whether trial options are accessible to patients from various socioeconomic backgrounds.

These use cases are fully described in the Appendix 1

Stage 6: Deploy & Monitor

Usefulness, Usability, and Efficacy: Stage 6

Re-assess the usability and effectiveness of the AI solution [25], [86]. Evaluate whether the usability or effectiveness of the AI solution has changed when deployed in an actual clinical environment.

Evaluate how the AI solution integrates in the workflow [12], [60], [61]. Determine whether a workflow integration assessment has been completed and documented. Upon deployment of the AI solution, assess whether it accounts for the flow of people and tasks throughout physical and digital environments. In addition, evaluate whether users are ignoring or developing work-arounds to deal with the solution, as well as whether the solution hinders patient-clinician interaction.

Assess the mechanisms in place to monitor AI solution performance over time [14], [61]. For this step, note whether there is a feedback loop for consistent detection of issues and defects, including a triage process to facilitate continuous improvement and monitoring. Also note whether there is a governance plan and documentation of responsibilities, and clarify whether the needs of accountable users are supported.

Consider the process for managing disagreement between end users and AI solution output [120]. Clarify whether a plan exists to manage clinician disagreements with AI solution output (e.g., human in the loop, human override, etc.).

Review processes around end user feedback to support continuous design [127]. Determine whether end user feedback has been solicited, considered, and used to continuously refine the AI solution.

Consider how the anticipated benefits, risks, and costs of the AI solution compare with the actual benefits, risks, and costs when used in the deployment environment [62]. Evaluate whether workflow error rates and response rates improve after the deployment of the AI solution. As part of this assessment, note whether the solution is superior to standard of care for a given outcome or outcomes, as well as whether the relative benefit is documented.

Assess the actions users take after interacting with the AI solution [121]. Determine whether the tasks involving use of the AI solution are adequately supported, and note whether these actions are different than originally anticipated.

Analyze user trust in the AI solution and consider how to support trust building [63], [64], [65]. For this step, the team should identify whether limitations (e.g., specific usage scenarios,

measures) for the intended use cases of the AI solution have been clearly expressed in nontechnical terms. In addition, determine whether users have access to relevant transparency and safety information (e.g., Model Cards).

Consider the patients/situations supported by the intended use of the AI solution, as well as tasks it does not support [56]. Determine whether there are clear inclusion/exclusion criteria for use of the AI solution. The team should also evaluate whether the solution is more or less useful for certain patient populations (e.g., pregnant women, low-risk patients, patients over age 50).

Fairness: Stage 6

Evaluate the potential for data drift monitoring methods to impact bias and fairness [14], [128], [129]. Determine whether data drift (inputs, outputs, and outcomes) in the AI system – along with impact on performance and clinical outcomes – will be monitored over the entire deployment population. In addition, evaluate whether monitoring protocols are in place for relevant sociodemographic subgroups in order to minimize unfair or systemic impacts. Finally, ascertain whether there are technical definitions for "significant" drift along with adequate justification for those definitions.

Identify persons responsible for effective monitoring of bias and fairness considerations for the AI solution [73]. It is important to identify the parties responsible for data monitoring and note whether they are qualified for those tasks. This includes verifying that responsible parties are able to access relevant social, ethical, legal, human factors and/or clinical stakeholders/advisers in the event that specific problems arise.

Consider whether adequate mitigation measures are in place to counter AI model drift [14], [130], [131], [132]. Assess whether the criterion defining "significant model drift" allows it to be detected before it impacts many people. Note whether automatic, easy-to-interpret notifications are provided to signal model drift on an ongoing basis to responsible individuals. If not, and if model performance drift must be manually evaluated, determine whether specific time intervals have been defined, along with adequate justification.

Consider if the impact of AI system bias is monitored effectively [133], [134]. Determine whether the AI system will be monitored to identify drift or bias and note timescales used to routinely evaluate AI system fairness. Also note whether data and model security are routinely monitored, and whether the timescales and frequency of security checks have been defined and adequately justified.

Consider how impacted populations can provide feedback and how associated procedures may impact bias and fairness [49], [135]. In case of an adverse event or large-scale issue, determine whether there is a clearly defined, unbiased (to business interests), and easy-to-access

way that affected individuals or groups can provide feedback or seek guidance, and whether these processes are consistent with state and federal policies. In addition, determine whether there is a way for impacted individuals to provide timely feedback about whether they were satisfied with the care/service that was provided with input from an AI solution, and whether it is equally accessible to all relevant subgroups. Finally, evaluate whether there are some individuals or groups who are systematically excluded from giving feedback due to language barriers or factors related to ability or access.

Evaluate the risks of model performance drift in the change from pilot to full deployment [134]. Identify whether there are long-term risks associated with the AI system or its performance (whether overall or by subgroup) that could not be measured during small-scale implementation, but that could be measured post-deployment. Assess whether short- and long-term impacts (and directions of impact) for AI system drifts have been considered.

Consider how drifts in AI system performance can impact bias, fairness, and the balanced allocation of resources, access, and/or outcomes [134]. Determine whether AI system performance and parity (inputs, outputs, and outcomes) will be monitored for significant drift across time for the entire population and for relevant sociodemographic subgroups to minimize unfair or systemic impacts. Determine whether specific criteria are defined for how "significant" shifts in model performance within subgroups (if differing from overall population due to sample size/error), or in parity between subgroups, should be defined, with adequate justification for the choice of criteria. This is consideration would apply to open and closed AI systems.

Clarify accountability for the effects of data/model breaches or data/AI system performance drift [73]. Regardless of whether an AI solution is purchased from a third-party vendor or developed internally, determine whether accountable parties for data/AI system breaches have been identified.

Evaluate whether transition from pilot to full deployment, changes in context, and/or changes in hardware/software alter the AI solution's performance [136], [137]. Determine whether AI system performance varies as a function of deployment site (rural vs. urban, community clinic vs. academic medical center, etc.) or deployment context (type of device or source of device/assay for specific input data, type of population most seen, etc.). Also assess whether data quality differs across various deployment sites and whether it affects the solution's performance. In addition, determine whether issues related to data quality affect AI system performance monitoring efforts or AI system performance in some subgroups more than others.

Consider how the transition to full deployment could affect the appropriate combination of human and automated decision-making with regard to fairness and bias [49], [138]. The team should once again assess whether end users will use output from the AI solution on its own to arrive at a decision or if that input will be used in tandem with other information. Also, determine whether the end user is able or required to override an AI system decision and evaluate whether the specific conditions for such an override, along with its potential impacts,

have been defined clearly in the full deployment context. In addition, determine whether there are differences in the pilot and deployment setting/processes that could alter how shared or automated decision-processes affect fairness and/or bias (e.g., is there less time? Is the population more heterogeneous? Is the patient flow greater?).

Evaluate whether and how affected groups will be informed about the role of an AI solution in providing their care [139], [140]. Determine whether there is an accessible, easy-to-understand way for affected groups to know that part of their treatment was determined by an AI solution and also note whether providing this information could increase the risk of harm or reduce the potential benefit. In addition, ascertain whether human factors or behavioral science experts have been consulted to determine how to present this information in a way that improves trust and patient agency.

Consider providing appropriate end-user feedback loops regarding bias and fairness related to use of the AI solution [141]. Determine whether end users or implementers can provide timely feedback about AI system performance, accuracy, or operational challenges that could increase harm or reduce benefit. In addition, assess whether feedback strategies are simple, informative, easy, and quick to access and complete, and whether feedback is reviewed in a timely manner (thereby preventing any existing issues from escalating/causing harm).

Safety and Reliability: Stage 6

Using a risk-based approach to patient safety, implement risk management and assessment methods from conception through to deployment of the AI solution [5], [12], [39], [49], [76], [77], [91]. Similar to Stage 5, risks to patient safety or observed instances of harm (including indirect harm) should be reported at a predetermined frequency to the developer and implementer organizations through a local governance process. These include issues, errors or malfunctions related to AI solution output, recommendations, and supporting software/hardware, including details about the rate of occurrence, apparent causes, whether the issues could be corrected and how, and any significant potential effects on patient care. A process should be established to report harms to the developer team, whether internal or external to the implementer organization, and the development team can share potential issues and expected risks and how those can be managed.

Trends and pattern detection of safety events for a given solution should be monitored. If there are safety issues and/or poor outcomes, thresholds should be established to decide whether the Al solution should continue to operate, needs to be refined, or whether it should be sunsetted.

Appropriate implementation of clearly structured and consistently repeatable decision-making processes by implementer organizations can assure that efforts to minimize patient safety risk and promote patient safety have been considered.

The implementer team should be aware of any effect that changes to architecture and code have to the level of risk, and there should be processes in place to manage risk arising from changes to system, environment, and data. Attention to detail is critical in areas of underlying implementation of the algorithm - a simple data overwrite can potentially lead to an adverse patient safety impact.

Finally, when possible, ensure that the AI solution is designed to keep a human in the loop to contest or override its output. When this is not possible, quality control processes can be implemented to assess accuracy and safety.

Maintain a monitoring process for adverse events (AEs) and serious adverse events (SAEs) [49]. For this step, monitor continuously for safety risks and their frequently, adhering to clear thresholds for detecting safety issues. Adverse events, including serious adverse events, should be reported and mitigated consistently, while classifying safety risks by severity and frequency. From a reporting perspective, when applicable, assess whether there is compliance with the organization's standard for "adverse event" and "serious adverse event, " and whether these are tracked separately and mitigated consistently. Finally, determine whether there are plans for sunsetting the AI solution (triggering a back-up plan) and initiating a safety investigation.

At frequent intervals throughout its deployment, review whether the AI solution has become outdated or no longer medically relevant [49], [77]. Determine the solution's clinical relevance (or input variables). In addition, clarify whether there are processes to identify incorrect or out-of-date knowledge/recommendations and to determine when a newer AI system should be used. Also, note whether there are processes for exploring whether there are better AI solutions for given tasks.

Maintain robust reporting and recall procedures for promptly notifying developer organizations, relevant agencies, and all users when safety concerns are discovered during implementation at multiple organizations [95], [126]. Flaws/biases/safety concerns related to AI solutions may be discovered during pilots and implementation by multiple organizations. The implementer team should consider how those are reported to the developer team so that the responsible organization, the FDA/other relevant agencies, and all AI solution users will be notified in a timely manner. In addition, identify how recalls/corrections will be managed and ascertain whether there is a defined process and threshold for reporting potential patient harm. Also, if safety concerns meet the defined risk threshold at one system, triggering a delay and re-evaluation, clarify whether this will be shared with the developer team and regulators, and determine whether there is a process in place for developer teams to share recalls and corrections with implementer teams.

Implement proper access controls and audit trail mechanisms consistent with the intended use of the AI solution [39]. Implementation of proper access controls and audit trail mechanisms should be balanced with the intended use of the AI solution. For this reason, the team should assess whether there are proper access controls in place for the AI solution,

including an audit trail that will allow an independent reviewer to determine who can access the interface. In addition, determine whether there is an audit trail that can determine how decisions are made based on the AI solution's output.

Ensure that unintended or unforeseen uses of the AI solution, similar to off-label use, can be reported [76], [93]. Determine whether there is a process to assess and report outcomes for unindicated or "off label" uses of AI solutions, as well as whether end users or other stakeholders are able to report such uses to local leadership. In addition, ascertain whether there are intermittent audits of how the AI solution is used vs. its intended purpose.

Conduct an impact analysis of the AI solution, analyzing safety and measures of benefit and bias, quantitatively and qualitatively, especially when the solution is updated [39]. Note whether an impact analysis plan is in place for any AI solution changes (e.g., for planned software updates), so that safety, effectiveness, and performance are not compromised. Also assess whether there is a process to ensure that all environment, application, and model updates to the AI solution are appropriately tested. Finally, determine whether version changes (and impact testing of version changes) are documented.

Implement measures to mitigate automation bias [42], [49], [92], [125]. *Automation bias* is the propensity of humans to over-rely on a suggestion from an automated system; this bias can potentially raise safety issues. In the context of CDS, automation bias can result in errors of commission (following incorrect advice) or omission (failing to act because of not being prompted to do so). For this reason, it is important to determine whether the potential for automation bias is measured for risk assessment and management purposes (for example, determining whether the incorrect AI system output can be detected and how it may have potential impact on subsequent decision making).

Ensure that AI solutions are labeled and updated with transparent information about their development, and disclose potential limitations including dataset, model, and system constraints, as well as important details about accuracy, error rates, generalizability, and clinical implications [42], [49], [91], [92], [93], [94]. Determine whether limitations, contraindications, accuracy, error rates, generalizability, robustness, reproducibility, clinical implications, and interpretation of results from the AI solution have been disclosed. Evaluate whether the implementer team incorporates an explanation to the clinician or end user regarding the reason(s) the solution made or suggested a particular decision. In addition, determine whether the developer team incorporates a method (e.g., Model Cards) to make it clear to end users that they are interacting with an AI system. Assess whether the AI system is labeled with a plain-language explanation of how the AI model was developed, its intended purpose, and its limitations and safety risks (e.g., type of model, description of dataset used to create the model, results from clinical studies, and which subpopulations were underrepresented in the training and test sets).

The team should also ensure that transparency information includes a clear explanation of the AI system's limitations and clinical implications (e.g., error rates, contraindications, generalizability, reproducibility, robustness, etc.). Along with that, there should be a process for updating information based on newly discovered limitations of local deployment in the implementer environment.

Ensure that users understand that no bug fixes, updates, patches, or technical support will be available once end-of-life (EOL) is signed off [39]. Technical support may include removal, migrating patients to a new AI solution, safe archival of user information, appropriately safeguarding patient data and any other confidential data, etc. It is important to clarify whether the implementer team's organization monitors and sunsets AI algorithms that are no longer supported and assess whether there is a plan for handling developer-driven EOL processes. It is also important to have an EOL plan for patient data, data migration, archival, etc. and clarify whether end users will receive support in the AI solution's absence (e.g., contact information, onboarding to a new solution, etc.).

Implement quality control techniques and technical standards to support supply chain risk management for AI solutions [142]. Technical standards assure purchasers and users that appropriate safety-focused measures are in place. Tools for trustworthy AI such as quality control techniques and technical standards can support supply chain risk management. These tools can also drive uptake and adoption of AI solutions by building justified trust in these systems by giving users confidence that key AI-related risks have been identified, addressed, and mitigated across the supply chain.

For solutions purchased from third-party vendors, the implementer team should engage in quality control techniques and/or technical standards to support the developer's supply chain risk management that include describing how safety risks should be reported. In addition, the team should ascertain whether the developer has provided a clear list regarding key AI system-related risks that have been identified, addressed, and mitigated across the supply chain or in other organizations. Also assess whether there is a description of measures that developer and implementer organizations should take to ensure the safety of AI solutions.

When updates are incorporated into an AI solution, conduct a rigorous impact analysis to ensure that safety and effectiveness are not compromised [39], [110]. For this step, determine whether there is a process or testing procedure in place to ensure that all updates and up-versioning of AI systems do not compromise patient safety. Ascertain whether different version changes (and any impact testing of version changes) are documented.

Transparency, Intelligibility, and Accountability: Stage 6

Report effectiveness to end users and key stakeholders [66]. Assess whether a preliminary study of AI solution effectiveness has been conducted and reported. Consider whether understanding of end users and key stakeholders can be measured by examining actions taken in

response to the AI solution and verifying the consistency of those actions against the defined limitations and intended use of the AI solution.

Consider whether patients are aware of the use of the AI solution, and whether it is necessary to communicate essential elements for understanding AI-based decisions [49]. It is important to establish a defined level of patient awareness regarding the AI solution's use as part of their care. If the end user is a healthcare provider, they should determine whether they are able to provide a satisfactory explanation of the AI solution output to a patient.

Maintain access to project-related and model information (e.g., clinical need, literature review, intended use, etc.) [81]. Determine whether there is a clearly defined method for patients and end users to access documentation about the AI solution and other project-related information. Also determine whether end users and patients will receive the same documentation (given various levels of expertise and health literacy).

Security and Privacy: Stage 6

Ensure that the impact of events is supported by incident response plans [3], [4]. Determine whether incident response plans are established, maintained, and tested according to policies for AI solutions. In addition, ensure that personnel responsible for AI system monitoring and incident response activities are trained on procedures to share incident impacts with stakeholders. Finally, assess whether personnel are responsible for reviewing, analyzing, and reporting incident impacts on the AI solution to stakeholders.

Ensure that impacted individuals and organizations are notified about cybersecurity incidents or privacy events [3], [4]. Determine whether a process aligned with legal, contractual, and regulatory requirements is in place to communicate to external stakeholders about cybersecurity incidents or privacy events.

Continuously evaluate privacy risk [3], [4]. Determine whether the implementer organization schedules privacy risk evaluations of its AI system environment according to an agreed upon interval (i.e., quarterly, annually, etc.). Key factors that can affect the degree of privacy risk include the organization's business environment (e.g., introduction of new technologies), governance (e.g., legal obligations, risk tolerance), data processing, and systems/products/services.

Verify that policies, processes, and procedures for assessing and communicating progress for compliance with legal requirements and privacy policies are in place [3], [4]. Assess whether the implementer organization has established policies and processes for communicating progress on compliance with legal requirements and managing privacy risk.

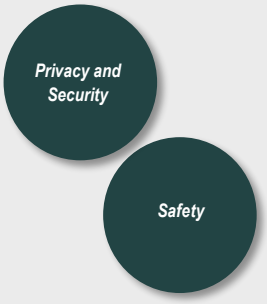
Use Case-Dependent Considerations for Stage 6

Use Case Dependent Considerations for Stage 6

Generative AI Use Case (EHR Query and Extraction)

Clear plans for updating and maintaining the AI system are crucial to ensure its ongoing effectiveness and security. Establishing a system for collecting and incorporating user feedback in real-time can aid in continuously improving usefulness, transparency, and identifying safety risks. Dynamic consent mechanisms, allowing patients to adjust their consent regarding data use over time and with emerging technologies or uses, can be considered as well.

Claims-Based Outpatient AI Use Case (Care Management)



Defining an "incident" within risk-stratification models for care management poses significant challenges. Discussion surrounding the appropriate level in the workflow for defining incidents, along with adequate justification, is necessary. Feedback loops play a crucial role in improving risk classification models. Providers and care managers often have comprehensive knowledge of patients over time, potentially surpassing the information available as input for AI models. Evidence suggests that care coordinators and providers may identify individuals as high risk despite not being identified as such by the algorithm, influenced by various factors.

Genomics AI Use Case (Precision Oncology with Genomic Markers)

The data powering the AI system must reflect current knowledge, necessitating continuous checks to determine the frequency of data updates for open clinical trials, treatment indications, guidelines, and biomarker-selection criteria. Genetic datasets often exhibit biases towards certain ethnicities and sample types, underscoring the importance of specifying the represented sub-group in the input data. Certain sub-groups are underrepresented in clinical trials for precision genomics in cancer, posing challenges for fair, inclusive, and accessible healthcare. AI systems can play a crucial role in addressing this disparity by measuring and providing information to increase participation. For instance, in cancer, race-specific variations in the occurrence and frequency of genomic aberrations have been observed [143]. However, existing datasets commonly used to train AI models, such as the Cancer Genome Atlas (TCGA), are predominantly composed of white individuals with European ancestry, indicating inherent biases [144]. Additionally, biases exist within these datasets, with limited representation of metastatic tumors [144]

These use cases are fully described in Appendix 1

8. Pathway for Continuous Learning

Artificial intelligence represents a paradigm shift in healthcare delivery. As AI solutions advance, it is imperative to ensure that they uphold the core principles laid out in this Guide while adapting to emerging patient needs in a rapidly evolving healthcare landscape. Continuous improvement is crucial to foster trust, proactively tackle challenges, seamlessly integrate AI solutions, and maintain alignment with ethical standards, patient rights, and user expectations. A culture of continuous improvement is fundamental for all stakeholders throughout the health AI lifecycle, as emphasized in this Guide.

A Learning Health System (LHS) provides a framework to support this culture of continuous improvement, systematically integrating internal experiences and external evidence. By leveraging LHS principles, organizations can innovate and continuously enhance their AI solutions based on data-driven insights, strategic priorities, and the evolving patient and end-user needs. Regular updates and feedback mechanisms are vital, ensuring that AI systems develop and improve with the latest clinical insights, medical guidelines, and patient feedback to enhance usability and efficacy. Through structured feedback mechanisms and quantitative analysis, the LHS framework effectively monitors AI systems for measures like error rates and changes in user satisfaction. Tools to detect and address model drift are crucial to ensure ongoing accuracy and relevance.

Governance within the LHS framework ensures that AI systems align with ethics and quality standards, empowering the health system’s mission to innovate while maintaining accountability. By adopting the principles and considerations outlined in the Guide, healthcare stakeholders not only embrace a new era of innovation but also commit to continuous improvement to meet evolving needs. The journey toward harnessing the full potential of AI is just beginning. Through collaboration and collective effort, stakeholders can pave the way for a dynamic era, ensuring that AI solutions not only enhance patient care but also uphold ethical standards and patient rights.

References

[1] L. Adams *et al.*, “Artificial Intelligence in Health, Health Care, and Biomedical Science: An AI Code of Conduct Principles and Commitments Discussion Draft,” *NAM Perspect.*, Apr. 2024, doi: 10.31478/202404a.

[2] “Blueprint for an AI Bill of Rights | OSTP,” The White House. Accessed: Apr. 29, 2024. [Online]. Available: https://www.whitehouse.gov/ostp/ai-bill-of-rights/

[3] National Institute of Standards and Technology, “The NIST Cybersecurity Framework (CSF) 2.0,” National Institute of Standards and Technology, Gaithersburg, MD, NIST CSWP 29, Feb. 2024. doi: 10.6028/NIST.CSWP.29.

[4] National Institute of Standards and Technology, “NIST PRIVACY FRAMEWORK: A TOOL FOR IMPROVING PRIVACY THROUGH ENTERPRISE RISK MANAGEMENT, VERSION 1.0,” National Institute of Standards and Technology, Gaithersburg, MD, NIST CSWP 01162020, Jan. 2020. doi: 10.6028/NIST.CSWP.01162020.

[5] E. Tabassi, “NIST Artificial Intelligence Risk Management Framework (AI RMF 1.0),” National Institute of Standards and Technology (U.S.), Gaithersburg, MD, NIST AI 100-1, Jan. 2023. doi: 10.6028/NIST.AI.100-1.

[6] “Healthcare Sector Cybersecurity Framework Implementation Guide 1.1”.

[7] W. Raynor, *International Dictionary of Artificial Intelligence*. Routledge, 2020.

[8] E. B. [D-T.-30 Rep. Johnson, “Text - H.R.6216 - 116th Congress (2019-2020): National Artificial Intelligence Initiative Act of 2020.” Accessed: May 03, 2024. [Online]. Available: https://www.congress.gov/bill/116th-congress/house-bill/6216/text

[9] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, “Dissecting racial bias in an algorithm used to manage the health of populations,” *Science*, vol. 366, no. 6464, pp. 447–453, Oct. 2019, doi: 10.1126/science.aax2342.

[10] S. Dutchen, “The Importance of Nuance | Harvard Medicine Magazine.” Accessed: Apr. 24, 2024. [Online]. Available: https://magazine.hms.harvard.edu/articles/importance-nuance

[11] Coalition for Health AI, “Blueprint for Trustworthy AI: Implementation Guidance and Assurance for Healthcare.” Accessed: Apr. 22, 2024. [Online]. Available: https://www.coalitionforhealthai.org/papers/blueprint-for-trustworthy-ai_V1.0.pdf

[12] B. Vasey *et al.*, “Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI,” *BMJ*, vol. 377, p. e070904, May 2022, doi: 10.1136/bmj-2022-070904.

[13] A. Downing and E. Perakslis, “Health advertising on Facebook: Privacy and policy considerations,” *Patterns*, vol. 3, no. 9, Sep. 2022, doi: 10.1016/j.patter.2022.100561.

[14] S. G. Finlayson *et al.*, “The Clinician and Dataset Shift in Artificial Intelligence,” *N. Engl. J. Med.*, vol. 385, no. 3, pp. 283–286, Jul. 2021, doi: 10.1056/NEJMc2104626.

[15] N. K. Corrêa *et al.*, “Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance,” *Patterns*, vol. 4, no. 10, Oct. 2023, doi: 10.1016/j.patter.2023.100857.

[16] T. L. C. Patient Experts and Community Leaders, “AI Rights for Patients.” Accessed: Apr. 22, 2024. [Online]. Available: https://lightcollective.org/wp-content/uploads/2024/03/Collective-Digital-Rights-For-Patients_v1.0.pdf

[17] A. D. Bedoya *et al.*, "A framework for the oversight and local deployment of safe and high-quality prediction models," *J. Am. Med. Inform. Assoc. JAMIA*, vol. 29, no. 9, pp. 1631–1636, May 2022, doi: 10.1093/jamia/ocac078.
[18] "key decision points," Health AI Partnership. Accessed: May 03, 2024. [Online]. Available: https://healthaipartnership.org/key-decisions-in-adopting-an-ai-solution
[19] A. Ferlitsch, "Making the machine: the machine learning lifecycle," Google Cloud Blog. Accessed: May 03, 2024. [Online]. Available: https://cloud.google.com/blog/products/ai-machine-learning/making-the-machine-the-machine-learning-lifecycle
[20] S. Kearney and R. G. Kenny, "SAS in Healthcare: Duke-SAS Health Innovation Lab." SAS Institute Inc.
[21] "Workbook | AI Toolkit." Accessed: May 03, 2024. [Online]. Available: https://www.ai-lawenforcement.org/assess/workbook
[22] K. E. Morse, S. C. Bagley, and N. H. Shah, "Estimate the hidden deployment cost of predictive models to improve patient care," *Nat. Med.*, vol. 26, no. 1, pp. 18–19, Jan. 2020, doi: 10.1038/s41591-019-0651-8.
[23] K. E. Henry *et al.*, "Human–machine teaming is key to AI adoption: clinicians' experiences with a deployed machine learning system," *Npj Digit. Med.*, vol. 5, no. 1, pp. 1–6, Jul. 2022, doi: 10.1038/s41746-022-00597-7.
[24] "ISO/IEC TS 5723:2022(en), Trustworthiness — Vocabulary." Accessed: Apr. 22, 2024. [Online]. Available: https://www.iso.org/obp/ui/#iso:std:iso-iec:ts:5723:ed-1:v1:en
[25] 14:00-17:00, "ISO 9241-11:2018," ISO. Accessed: Apr. 22, 2024. [Online]. Available: https://www.iso.org/standard/63500.html
[26] S. Corbett-Davies, J. D. Gaebler, H. Nilforoshan, R. Shroff, and S. Goel, "The Measure and Mismeasure of Fairness," 2018, doi: 10.48550/ARXIV.1808.00023.
[27] H. E. Wang *et al.*, "A bias evaluation checklist for predictive models and its pilot application for 30-day hospital readmission models," *J. Am. Med. Inform. Assoc. JAMIA*, vol. 29, no. 8, pp. 1323–1333, Jul. 2022, doi: 10.1093/jamia/ocac065.
[28] P. Braveman and S. Gruskin, "Defining equity in health," *J. Epidemiol. Community Health*, vol. 57, no. 4, pp. 254–258, Apr. 2003, doi: 10.1136/jech.57.4.254.
[29] T. Li, D. Khashabi, T. Khot, A. Sabharwal, and V. Srikumar, "UNQOVERing Stereotyping Biases via Underspecified Questions," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, T. Cohn, Y. He, and Y. Liu, Eds., Online: Association for Computational Linguistics, Nov. 2020, pp. 3475–3489. doi: 10.18653/v1/2020.findings-emnlp.311.
[30] "Information bias," Catalog of Bias. Accessed: Apr. 22, 2024. [Online]. Available: https://catalogofbias.org/biases/information-bias/
[31] T. Panch, H. Mattie, and R. Atun, "Artificial intelligence and algorithmic bias: implications for health systems," *J. Glob. Health*, vol. 9, no. 2, p. 010318, Dec. 2019, doi: 10.7189/jogh.09.020318.
[32] R. R. Bond *et al.*, "Automation bias in medicine: The influence of automated diagnoses on interpreter accuracy and uncertainty when reading electrocardiograms," *J. Electrocardiol.*, vol. 51, no. 6, pp. S6–S11, Nov. 2018, doi: 10.1016/j.jelectrocard.2018.08.007.
[33] *To Err Is Human: Building a Safer Health System*. Washington, D.C.: National Academies Press, 2000, p. 9728. doi: 10.17226/9728.

[34] J. B. Cooper, D. M. Gaba, B. Liang, D. Woods, and L. N. Blum, “The National Patient Safety Foundation agenda for research and development in patient safety,” *MedGenMed Medscape Gen. Med.*, vol. 2, no. 3, p. E38, Jul. 2000.

[35] A. Subbaswamy and S. Saria, “From development to deployment: dataset shift, causality, and shift-stable models in health AI,” *Biostat. Oxf. Engl.*, vol. 21, no. 2, pp. 345–352, Apr. 2020, doi: 10.1093/biostatistics/kxz041.

[36] “Health Data, Technology, and Interoperability: Certification Program Updates, Algorithm Transparency, and Information Sharing,” Federal Register. Accessed: Apr. 22, 2024. [Online]. Available: https://www.federalregister.gov/documents/2024/01/09/2023-28857/health-data-technology-and-interoperability-certification-program-updates-algorithm-transparency-and

[37] NIST, “The Language of Trustworthy AI: An In-Depth Glossary of Terms (March 22, 2023) - Google Drive.” Accessed: May 03, 2024. [Online]. Available: https://docs.google.com/spreadsheets/d/e/2PACX-1vTRBYglcOtgaMrdF11aFxfEY3EmB31zslYI4q2_7ZZ8z_1lKm7OHtF0t4xIsckuogNZ3hRZAaDQuv_K/pubhtml

[38] N. J. Economou-Zavlanos *et al.*, “Translating ethical and quality principles for the effective, safe and fair development, deployment and use of artificial intelligence technologies in healthcare,” *J. Am. Med. Inform. Assoc.*, vol. 31, no. 3, pp. 705–713, Mar. 2024, doi: 10.1093/jamia/ocad221.

[39] “IMDRF Proposed Document: Software as a Medical Device (SaMD): Clinical Evaluation.” Aug. 05, 2016. [Online]. Available: https://www.imdrf.org/sites/default/files/2021-09/imdrf-cons-samd-ce.pdf

[40] L. L. Novak *et al.*, “Clinical use of artificial intelligence requires AI-capable organizations,” *JAMIA Open*, vol. 6, no. 2, p. ooad028, Jul. 2023, doi: 10.1093/jamiaopen/ooad028.

[41] K. L. Dempsey *et al.*, “Information Security Continuous Monitoring (ISCM) for federal information systems and organizations,” National Institute of Standards and Technology, Gaithersburg, MD, NIST SP 800-137, 2011. doi: 10.6028/NIST.SP.800-137.

[42] European Commission, “Ethics by Design and Ethics of Use Approaches for Artificial Intelligence.” Accessed: Apr. 22, 2024. [Online]. Available: https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/ethics-by-design-and-ethics-of-use-approaches-for-artificial-intelligence_he_en.pdf

[43] “Safe and Effective Systems, AI Bill of Rights,” The White House. Accessed: Apr. 22, 2024. [Online]. Available: https://www.whitehouse.gov/ostp/ai-bill-of-rights/safe-and-effective-systems-3/

[44] S. Reddy, S. Allan, S. Coghlan, and P. Cooper, “A governance model for the application of AI in health care,” *J. Am. Med. Inform. Assoc. JAMIA*, vol. 27, no. 3, pp. 491–497, Nov. 2019, doi: 10.1093/jamia/ocz192.

[45] J. Y. Kim *et al.*, “Development and preliminary testing of Health Equity Across the AI Lifecycle (HEAAL): A framework for healthcare delivery organizations to mitigate the risk of AI solutions worsening health inequities.” medRxiv, p. 2023.10.16.23297076, Feb. 19, 2024. doi: 10.1101/2023.10.16.23297076.

[46] D. D. Silva and D. Alahakoon, “An artificial intelligence life cycle: From conception to production,” *Patterns*, vol. 3, no. 6, Jun. 2022, doi: 10.1016/j.patter.2022.100489.

[47] D. Oniani *et al.*, “Adopting and expanding ethical principles for generative artificial intelligence from military to healthcare,” *Npj Digit. Med.*, vol. 6, no. 1, pp. 1–10, Dec. 2023, doi: 10.1038/s41746-023-00965-x.
[48] E. O. Nsoesie and S. Galea, “Towards better Data Science to address racial bias and health equity,” *PNAS Nexus*, vol. 1, no. 3, p. pgac120, Jul. 2022, doi: 10.1093/pnasnexus/pgac120.
[49] “CHAI Workgroup Discussions (Apr-Dec 2023).” 2023.
[50] “Software as a Medical Device (SAMD): Clinical Evaluation - Guidance for Industry and Food and Drug Administration Staff,” 2017.
[51] N. H. Shah *et al.*, “A Nationwide Network of Health AI Assurance Laboratories,” *JAMA*, vol. 331, no. 3, pp. 245–249, Jan. 2024, doi: 10.1001/jama.2023.26930.
[52] A. Wong *et al.*, “External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients,” *JAMA Intern. Med.*, vol. 181, no. 8, pp. 1065–1070, Aug. 2021, doi: 10.1001/jamainternmed.2021.2626.
[53] “Guidance for Industry: Q8(R2) Pharmaceutical Development.” U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER), Nov. 200AD. Accessed: May 03, 2024. [Online]. Available: https://www.fda.gov/media/71535/download
[54] “Computer Software Assurance for Production and Quality System Software: Draft Guidance for Industry and FDA Staff.” U.S. Food and Drug Administration, Sep. 13, 2022.
[55] K. Jung *et al.*, “A framework for making predictive models useful in practice,” *J. Am. Med. Inform. Assoc.*, vol. 28, no. 6, pp. 1149–1158, Jun. 2021, doi: 10.1093/jamia/ocaa318.
[56] R. C. Li, S. M. Asch, and N. H. Shah, “Developing a delivery science for artificial intelligence in healthcare,” *Npj Digit. Med.*, vol. 3, no. 1, pp. 1–3, Aug. 2020, doi: 10.1038/s41746-020-00318-y.
[57] M. G. Seneviratne *et al.*, “User-centred design for machine learning in health care: a case study from care management,” *BMJ Health Care Inform.*, vol. 29, no. 1, p. e100656, Oct. 2022, doi: 10.1136/bmjhci-2022-100656.
[58] J. Wiens *et al.*, “Do no harm: a roadmap for responsible machine learning for health care,” *Nat. Med.*, vol. 25, no. 9, pp. 1337–1340, Sep. 2019, doi: 10.1038/s41591-019-0548-6.
[59] M. Wornow, E. Gyang Ross, A. Callahan, and N. H. Shah, “APLUS: A Python library for usefulness simulations of machine learning models in healthcare,” *J. Biomed. Inform.*, vol. 139, p. 104319, Mar. 2023, doi: 10.1016/j.jbi.2023.104319.
[60] M. E. Salwei *et al.*, “Workflow integration analysis of a human factors-based clinical decision support in the emergency department,” *Appl. Ergon.*, vol. 97, p. 103498, Nov. 2021, doi: 10.1016/j.apergo.2021.103498.
[61] M. E. Salwei and P. Carayon, “A Sociotechnical Systems Framework for the Application of Artificial Intelligence in Health Care Delivery,” *J. Cogn. Eng. Decis. Mak.*, vol. 16, no. 4, pp. 194–206, Dec. 2022, doi: 10.1177/15553434221097357.
[62] Y. Park, G. P. Jackson, M. A. Foreman, D. Gruen, J. Hu, and A. K. Das, “Evaluating artificial intelligence in medicine: phases of clinical research,” *JAMIA Open*, vol. 3, no. 3, pp. 326–331, Oct. 2020, doi: 10.1093/jamiaopen/ooaa033.

[63] N. C. Benda, L. L. Novak, C. Reale, and J. S. Ancker, "Trust in AI: why we should be designing for APPROPRIATE reliance," *J. Am. Med. Inform. Assoc.*, vol. 29, no. 1, pp. 207–212, Jan. 2022, doi: 10.1093/jamia/ocab238.

[64] O. Asan, A. E. Bayrak, and A. Choudhury, "Artificial Intelligence and Human Trust in Healthcare: Focus on Clinicians," *J. Med. Internet Res.*, vol. 22, no. 6, p. e15154, Jun. 2020, doi: 10.2196/15154.

[65] R. Hoffman, S. Mueller, G. Klein, and J. Litman, "Measuring Trust in the XAI Context." OSF, Nov. 01, 2021. doi: 10.31234/osf.io/e3kv9.

[66] X. Liu *et al.*, "Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension," *Lancet Digit. Health*, vol. 2, no. 10, pp. e537–e548, Oct. 2020, doi: 10.1016/S2589-7500(20)30218-1.

[67] J. Gallifant *et al.*, "Disparity dashboards: an evaluation of the literature and framework for health equity improvement," *Lancet Digit. Health*, vol. 5, no. 11, pp. e831–e839, Nov. 2023, doi: 10.1016/S2589-7500(23)00150-4.

[68] A. Rajkomar, M. Hardt, M. D. Howell, G. Corrado, and M. H. Chin, "Ensuring Fairness in Machine Learning to Advance Health Equity," *Ann. Intern. Med.*, vol. 169, no. 12, pp. 866–872, Dec. 2018, doi: 10.7326/M18-1990.

[69] H. Suresh and J. Guttag, "A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle," in *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, in EAAMO '21. New York, NY, USA: Association for Computing Machinery, Nov. 2021, pp. 1–9. doi: 10.1145/3465416.3483305.

[70] D. S. Char, M. D. Abràmoff, and C. Feudtner, "Identifying Ethical Considerations for Machine Learning Healthcare Applications," *Am. J. Bioeth. AJOB*, vol. 20, no. 11, pp. 7–17, Nov. 2020, doi: 10.1080/15265161.2020.1819469.

[71] P. Omoumi *et al.*, "To buy or not to buy-evaluating commercial AI solutions in radiology (the ECLAIR guidelines)," *Eur. Radiol.*, vol. 31, no. 6, pp. 3786–3796, Jun. 2021, doi: 10.1007/s00330-020-07684-x.

[72] S. Tatineni, "Ethical Considerations in AI and Data Science: Bias, Fairness, and Accountability".

[73] R. Schwartz, A. Vassilev, K. Greene, L. Perine, A. Burt, and P. Hall, "Towards a standard for identifying and managing bias in artificial intelligence," National Institute of Standards and Technology (U.S.), Gaithersburg, MD, NIST SP 1270, Mar. 2022. doi: 10.6028/NIST.SP.1270.

[74] M. Portela, "Towards a meaningful human oversight of automated decision-making systems," Digital Future Society. Accessed: Apr. 22, 2024. [Online]. Available: https://digitalfuturesociety.com/report/towards-a-meaningful-human-oversight-of-automated-decision-making-systems/

[75] D. Sele and M. Chugunova, "Putting a Human in the Loop: Increasing Uptake, but Decreasing Accuracy of Automated Decision-Making." Rochester, NY, Nov. 18, 2022. Accessed: Apr. 22, 2024. [Online]. Available: https://papers.ssrn.com/abstract=4285645

[76] "NTIA Artificial Intelligence Accountability Policy Report MARCH 2024".

[77] D. Vanderpool, "The Standard of Care," *Innov. Clin. Neurosci.*, vol. 18, no. 7–9, pp. 50–51, 2021.

[78] A. Coston, A. Kawakami, H. Zhu, K. Holstein, and H. Heidari, “A Validity Perspective on Evaluating the Justified Use of Data-driven Decision-making Algorithms.” arXiv, Feb. 14, 2023. doi: 10.48550/arXiv.2206.14983.

[79] N. Martinez-Martin *et al.*, “Ethical issues in using ambient intelligence in health-care settings,” *Lancet Digit. Health*, vol. 3, no. 2, pp. e115–e123, Feb. 2021, doi: 10.1016/S2589-7500(20)30275-2.

[80] K. G. M. Moons *et al.*, “Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration,” *Ann. Intern. Med.*, vol. 162, no. 1, pp. W1-73, Jan. 2015, doi: 10.7326/M14-0698.

[81] M. Mitchell *et al.*, “Model Cards for Model Reporting,” in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, in FAT* ’19. New York, NY, USA: Association for Computing Machinery, Jan. 2019, pp. 220–229. doi: 10.1145/3287560.3287596.

[82] S. Cruz Rivera, X. Liu, A.-W. Chan, A. K. Denniston, and M. J. Calvert, “Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension,” *Nat. Med.*, vol. 26, no. 9, pp. 1351–1363, Sep. 2020, doi: 10.1038/s41591-020-1037-7.

[83] T. A. Brereton, M. M. Malik, M. Lifson, J. D. Greenwood, K. J. Peterson, and S. M. Overgaard, “The Role of Artificial Intelligence Model Documentation in Translational Science: Scoping Review,” *Interact. J. Med. Res.*, vol. 12, no. 1, p. e45903, Jul. 2023, doi: 10.2196/45903.

[84] E. Breck, S. Cai, E. Nielsen, M. Salib, and D. Sculley, “The ML test score: A rubric for ML production readiness and technical debt reduction,” in *2017 IEEE International Conference on Big Data (Big Data)*, Dec. 2017, pp. 1123–1132. doi: 10.1109/BigData.2017.8258038.

[85] “CFR - Code of Federal Regulations Title 21.” Accessed: Apr. 22, 2024. [Online]. Available: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm?fr=56.115

[86] 14:00-17:00, “ISO/TR 16982:2002,” ISO. Accessed: Apr. 22, 2024. [Online]. Available: https://www.iso.org/standard/31176.html

[87] J. Yin, K. Y. Ngiam, and H. H. Teo, “Role of Artificial Intelligence Applications in Real-Life Clinical Practice: Systematic Review,” *J. Med. Internet Res.*, vol. 23, no. 4, p. e25759, Apr. 2021, doi: 10.2196/25759.

[88] A. Kale, T. Nguyen, F. C. Harris, C. Li, J. Zhang, and X. Ma, “Provenance documentation to enable explainable and trustworthy AI: A literature review,” *Data Intell.*, vol. 5, no. 1, pp. 139–162, Mar. 2023, doi: 10.1162/dint_a_00119.

[89] N. Ivers *et al.*, “Audit and feedback: effects on professional practice and healthcare outcomes,” *Cochrane Database Syst. Rev.*, no. 6, p. CD000259, Jun. 2012, doi: 10.1002/14651858.CD000259.pub3.

[90] J. Shuldiner *et al.*, “Developing an Audit and Feedback Dashboard for Family Physicians: User-Centered Design Process,” *JMIR Hum. Factors*, vol. 10, p. e47718, Nov. 2023, doi: 10.2196/47718.

[91] “Google Responsible AI Practices,” Google AI. Accessed: Apr. 22, 2024. [Online]. Available: https://ai.google/responsibility/responsible-ai-practices/

[92] Food and Drug Administration, “Clinical Decision Support Software - Guidance for Industry and Food and Drug Administration Staff,” Sep. 2022.

[93] "TRIPOD-AI_round_1_summary.pdf," Dec. 2021, Accessed: Apr. 22, 2024. [Online]. Available: https://osf.io/https://osf.io/nskme
[94] M. P. Sendak, M. Gao, N. Brajer, and S. Balu, "Presenting machine learning model information to clinical end users with model facts labels," *Npj Digit. Med.*, vol. 3, no. 1, pp. 1–4, Mar. 2020, doi: 10.1038/s41746-020-0253-3.
[95] Food and Drug Administration, "Recalls, Corrections and Removals (Devices)," FDA. Accessed: Apr. 22, 2024. [Online]. Available: https://www.fda.gov/medical-devices/postmarket-requirements-devices/recalls-corrections-and-removals-devices
[96] M. Bayati *et al.*, "Data-Driven Decisions for Reducing Readmissions for Heart Failure: General Methodology and Case Study," *PLoS ONE*, vol. 9, no. 10, p. e109264, Oct. 2014, doi: 10.1371/journal.pone.0109264.
[97] W. Liang *et al.*, "Advances, challenges and opportunities in creating data for trustworthy AI," *Nat. Mach. Intell.*, vol. 4, no. 8, pp. 669–677, Aug. 2022, doi: 10.1038/s42256-022-00516-1.
[98] Y. Chen, E. W. Clayton, L. L. Novak, S. Anders, and B. Malin, "Human-Centered Design to Address Biases in Artificial Intelligence," *J. Med. Internet Res.*, vol. 25, no. 1, p. e43251, Mar. 2023, doi: 10.2196/43251.
[99] C. Meng, L. Trinh, N. Xu, J. Enouen, and Y. Liu, "Interpretability and fairness evaluation of deep learning models on MIMIC-IV dataset," *Sci. Rep.*, vol. 12, no. 1, p. 7166, May 2022, doi: 10.1038/s41598-022-11012-2.
[100] J. P. Cohen *et al.*, "Problems in the deployment of machine-learned models in health care," *CMAJ Can. Med. Assoc. J.*, vol. 193, no. 35, pp. E1391–E1394, Sep. 2021, doi: 10.1503/cmaj.202066.
[101] N. Zhou, Z. Zhang, V. N. Nair, H. Singhal, J. Chen, and A. Sudjianto, "Bias, Fairness, and Accountability with AI and ML Algorithms." arXiv, May 13, 2021. doi: 10.48550/arXiv.2105.06558.
[102] I. Y. Chen, E. Pierson, S. Rose, S. Joshi, K. Ferryman, and M. Ghassemi, "Ethical Machine Learning in Healthcare," *Annu. Rev. Biomed. Data Sci.*, vol. 4, no. Volume 4, 2021, pp. 123–144, Jul. 2021, doi: 10.1146/annurev-biodatasci-092820-114757.
[103] Y. J. Juhn *et al.*, "Assessing socioeconomic bias in machine learning algorithms in health care: a case study of the HOUSES index," *J. Am. Med. Inform. Assoc. JAMIA*, vol. 29, no. 7, pp. 1142–1151, Jun. 2022, doi: 10.1093/jamia/ocac052.
[104] Obermeyer, Ziad, Rebecca Nissan, Michael Stern, Stephanie Eaneff, Emily Joy Bembeneck, Sendhil Mullainathan, "Algorithmic Bias Playbook." Accessed: Apr. 22, 2024. [Online]. Available: https://www.ftc.gov/system/files/documents/public_events/1582978/algorithmic-bias-playbook.pdf
[105] A. de Hond *et al.*, "Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review," *Npj Digit. Med.*, vol. 5, Dec. 2022, doi: 10.1038/s41746-021-00549-7.
[106] S. Bozkurt *et al.*, "Reporting of demographic data and representativeness in machine learning models using electronic health records," *J. Am. Med. Inform. Assoc. JAMIA*, vol. 27, no. 12, pp. 1878–1884, Dec. 2020, doi: 10.1093/jamia/ocaa164.
[107] R. J. Chen *et al.*, "Algorithmic fairness in artificial intelligence for medicine and healthcare," *Nat. Biomed. Eng.*, vol. 7, no. 6, pp. 719–742, Jun. 2023, doi: 10.1038/s41551-023-01056-8.

[108] M. Pruski, “What does it mean for a clinical AI to be just: conflicts between local fairness and being fit-for-purpose?,” *J. Med. Ethics*, Feb. 2024, doi: 10.1136/jme-2023-109675.
[109] M. Matheny, S. T. Israni, M. Ahmed, and D. Whicher, “Artificial Intelligence in Health Care: The Hope, the Hype, the Promise, the Peril”.
[110] S. Saria, “Not All AI Is Created Equal: Strategies for Safe and Effective Adoption,” 2022.
[111] K. V. Iserson, “Informed consent for artificial intelligence in emergency medicine: A practical guide,” *Am. J. Emerg. Med.*, vol. 76, pp. 225–230, Feb. 2024, doi: 10.1016/j.ajem.2023.11.022.
[112] E. Tal, “Target specification bias, counterfactual prediction, and algorithmic fairness in healthcare,” in *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, in AIES ’23. New York, NY, USA: Association for Computing Machinery, Aug. 2023, pp. 312–321. doi: 10.1145/3600211.3604678.
[113] D. W. Bates, A. Auerbach, P. Schulam, A. Wright, and S. Saria, “Reporting and Implementing Interventions Involving Machine Learning and Artificial Intelligence,” *Ann. Intern. Med.*, vol. 172, no. 11 Suppl, pp. S137–S144, Jun. 2020, doi: 10.7326/M19-0872.
[114] O. Brown, A. Curtis, and J. Goodwin, “Principles for Evaluation of AI/ML Model Performance and Robustness.” arXiv, Jul. 06, 2021. doi: 10.48550/arXiv.2107.02868.
[115] E. A. M. Stanley, M. Wilms, and N. D. Forkert, “Disproportionate Subgroup Impacts and Other Challenges of Fairness in Artificial Intelligence for Medical Image Analysis,” in *Ethical and Philosophical Issues in Medical Imaging, Multimodal Learning and Fusion Across Scales for Clinical Decision Support, and Topological Data Analysis for Biomedical Imaging*, J. S. H. Baxter, I. Rekik, R. Eagleson, L. Zhou, T. Syeda-Mahmood, H. Wang, and M. Hajij, Eds., Cham: Springer Nature Switzerland, 2022, pp. 14–25. doi: 10.1007/978-3-031-23223-7_2.
[116] F. C. for B. E. and R. FDA Center for Devices and Radiological Health, “General Principles of Software Validation.” Accessed: Apr. 23, 2024. [Online]. Available: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/general-principles-software-validation
[117] A. Kiderman, U. Ilan, I. Gur, T. Bdolah-Abram, and M. Brezis, “Unexplained complaints in primary care: evidence of action bias,” *J. Fam. Pract.*, vol. 62, no. 8, pp. 408–413, Aug. 2013.
[118] S. M. Marx *et al.*, “Communication and mental processes: Experiential and analytic processing of uncertain climate information,” *Glob. Environ. Change*, vol. 17, no. 1, pp. 47–58, Feb. 2007, doi: 10.1016/j.gloenvcha.2006.10.004.
[119] A. Reddy *et al.*, “Risk Stratification Methods and Provision of Care Management Services in Comprehensive Primary Care Initiative Practices,” *Ann. Fam. Med.*, vol. 15, no. 5, pp. 451–454, Sep. 2017, doi: 10.1370/afm.2124.
[120] T. Grote and P. Berens, “How competitors become collaborators-Bridging the gap(s) between machine learning algorithms and clinicians,” *Bioethics*, vol. 36, no. 2, pp. 134–142, Feb. 2022, doi: 10.1111/bioe.12957.
[121] B. Shneiderman, “Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy,” *Int. J. Human–Computer Interact.*, vol. 36, no. 6, pp. 495–504, Apr. 2020, doi: 10.1080/10447318.2020.1741118.
[122] M. R. Endsley, “Supporting Human-AI Teams:Transparency, explainability, and situation awareness,” *Comput. Hum. Behav.*, vol. 140, p. 107574, Mar. 2023, doi: 10.1016/j.chb.2022.107574.

[123] A. M. Gustavson *et al.*, “Strategies to Bridge Equitable Implementation of Telehealth,” *Interact. J. Med. Res.*, vol. 12, p. e40358, May 2023, doi: 10.2196/40358.
[124] J. C. C. Kwong *et al.*, “The silent trial - the bridge between bench-to-bedside clinical AI applications,” *Front. Digit. Health*, vol. 4, p. 929508, Aug. 2022, doi: 10.3389/fdgth.2022.929508.
[125] T. Dratsch *et al.*, “Automation Bias in Mammography: The Impact of Artificial Intelligence BI-RADS Suggestions on Reader Performance,” *Radiology*, vol. 307, no. 4, p. e222176, May 2023, doi: 10.1148/radiol.222176.
[126] Food and Drug Administration, “Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence/Machine Learning (AI/ML)-Enabled Device Software Functions”.
[127] P. Carayon and M. E. Salwei, “Moving toward a sociotechnical systems approach to continuous health information technology design: the path forward for improving electronic health record usability and reducing clinician burnout,” *J. Am. Med. Inform. Assoc. JAMIA*, vol. 28, no. 5, pp. 1026–1028, Feb. 2021, doi: 10.1093/jamia/ocab002.
[128] A. Khoshravan Azar, B. Draghi, Y. Rotalinti, P. Myles, and A. Tucker, “The Impact of Bias on Drift Detection in AI Health Software,” in *Artificial Intelligence in Medicine*, J. M. Juarez, M. Marcos, G. Stiglic, and A. Tucker, Eds., Cham: Springer Nature Switzerland, 2023, pp. 313–322. doi: 10.1007/978-3-031-34344-5_37.
[129] K. Zadorozhny, P. Thoral, P. Elbers, and G. Cinà, “Out-of-Distribution Detection for Medical Applications: Guidelines for Practical Evaluation,” in *Multimodal AI in Healthcare: A Paradigm Shift in Health Intelligence*, A. Shaban-Nejad, M. Michalowski, and S. Bianco, Eds., Cham: Springer International Publishing, 2023, pp. 137–153. doi: 10.1007/978-3-031-14771-5_10.
[130] C. McLeod, R. Norman, E. Litton, B. R. Saville, S. Webb, and T. L. Snelling, “Choosing primary endpoints for clinical trials of health care interventions,” *Contemp. Clin. Trials Commun.*, vol. 16, p. 100486, Dec. 2019, doi: 10.1016/j.conctc.2019.100486.
[131] A. Kore *et al.*, “Empirical data drift detection experiments on real-world medical imaging data,” *Nat. Commun.*, vol. 15, no. 1, p. 1887, Feb. 2024, doi: 10.1038/s41467-024-46142-w.
[132] B. Sahiner, W. Chen, R. K. Samala, and N. Petrick, “Data drift in medical machine learning: implications and potential remedies,” *Br. J. Radiol.*, vol. 96, no. 1150, p. 20220878, Oct. 2023, doi: 10.1259/bjr.20220878.
[133] J. Feng *et al.*, “Clinical artificial intelligence quality improvement: towards continual monitoring and updating of AI algorithms in healthcare,” *Npj Digit. Med.*, vol. 5, no. 1, pp. 1–9, May 2022, doi: 10.1038/s41746-022-00611-y.
[134] J. Feng *et al.*, “Monitoring the performance of machine learning algorithms that induce feedback loops: what is the causal estimand?” arXiv, Feb. 26, 2024. Accessed: Apr. 23, 2024. [Online]. Available: http://arxiv.org/abs/2311.11463
[135] PatientEngagementHIT, “Patient Trust in Healthcare AI Relies on Use Case, But Familiarity Is Lacking,” PatientEngagementHIT. Accessed: Apr. 23, 2024. [Online]. Available: https://patientengagementhit.com/news/patient-trust-in-healthcare-ai-relies-on-use-case-but-familiarity-is-lacking
[136] V. S. M. Valbuena *et al.*, “Racial bias and reproducibility in pulse oximetry among medical and surgical inpatients in general care in the Veterans Health Administration

2013-19: multicenter, retrospective cohort study," *BMJ*, p. e069775, Jul. 2022, doi: 10.1136/bmj-2021-069775.
[137] J. Futoma, M. Simons, T. Panch, F. Doshi-Velez, and L. A. Celi, "The myth of generalisability in clinical research and machine learning in health care," *Lancet Digit. Health*, vol. 2, no. 9, pp. e489–e492, Sep. 2020, doi: 10.1016/S2589-7500(20)30186-2.
[138] U. Ehsan, Q. V. Liao, M. Muller, M. O. Riedl, and J. D. Weisz, "Expanding Explainability: Towards Social Transparency in AI systems," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, in CHI '21. New York, NY, USA: Association for Computing Machinery, May 2021, pp. 1–19. doi: 10.1145/3411764.3445188.
[139] A. Kiseleva, D. Kotzinos, and P. De Hert, "Transparency of AI in Healthcare as a Multilayered System of Accountabilities: Between Legal Requirements and Technical Limitations," *Front. Artif. Intell.*, vol. 5, May 2022, doi: 10.3389/frai.2022.879603.
[140] Z. Zhang, Y. Genc, D. Wang, M. E. Ahsen, and X. Fan, "Effect of AI Explanations on Human Perceptions of Patient-Facing AI-Powered Healthcare Systems," *J. Med. Syst.*, vol. 45, no. 6, p. 64, May 2021, doi: 10.1007/s10916-021-01743-6.
[141] S. I. Lambert *et al.*, "An integrative review on the acceptance of artificial intelligence among healthcare professionals in hospitals," *Npj Digit. Med.*, vol. 6, no. 1, pp. 1–14, Jun. 2023, doi: 10.1038/s41746-023-00852-5.
[142] UK Secretary of State for Science, Innovation and Technology, "A pro-innovation approach to AI regulation," GOV.UK. Accessed: Apr. 23, 2024. [Online]. Available: https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper
[143] B. Bhinder, C. Gilvary, N. S. Madhukar, and O. Elemento, "Artificial Intelligence in Cancer Research and Precision Medicine," *Cancer Discov.*, vol. 11, no. 4, pp. 900–915, Apr. 2021, doi: 10.1158/2159-8290.CD-21-0090.
[144] J. Yuan *et al.*, "Integrated Analysis of Genetic Ancestry and Genomic Alterations across Cancers," *Cancer Cell*, vol. 34, no. 4, pp. 549-560.e9, Oct. 2018, doi: 10.1016/j.ccell.2018.08.019.
[145] Y. Juhn and H. Liu, "Artificial intelligence approaches using natural language processing to advance EHR-based clinical research," *J. Allergy Clin. Immunol.*, vol. 145, no. 2, pp. 463–469, Feb. 2020, doi: 10.1016/j.jaci.2019.12.897.
[146] S. T. Wu *et al.*, "Automated chart review for asthma cohort identification using natural language processing: an exploratory study," *Ann. Allergy Asthma Immunol. Off. Publ. Am. Coll. Allergy Asthma Immunol.*, vol. 111, no. 5, pp. 364–369, Nov. 2013, doi: 10.1016/j.anai.2013.07.022.
[147] "A common type system for clinical natural language processing - PubMed." Accessed: Jun. 24, 2024. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/23286462/
[148] C. F. De Vries *et al.*, "Impact of Different Mammography Systems on Artificial Intelligence Performance in Breast Cancer Screening," *Radiol. Artif. Intell.*, vol. 5, no. 3, p. e220146, May 2023, doi: 10.1148/ryai.220146.
[149] K. Lång *et al.*, "Artificial intelligence-supported screen reading versus standard double reading in the Mammography Screening with Artificial Intelligence trial (MASAI): a clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded, screening accuracy study," *Lancet Oncol.*, vol. 24, no. 8, pp. 936–944, Aug. 2023, doi: 10.1016/S1470-2045(23)00298-X.

[150] C. D. Lehman and E. J. Topol, “Readiness for mammography and artificial intelligence,” *Lancet Lond. Engl.*, vol. 398, no. 10314, p. 1867, Nov. 2021, doi: 10.1016/S0140-6736(21)02484-3.
[151] P. Lewis *et al.*, “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.” arXiv, Apr. 12, 2021. doi: 10.48550/arXiv.2005.11401.
[152] T. Shen *et al.*, “Large Language Model Alignment: A Survey.” arXiv, Sep. 26, 2023. Accessed: Jun. 24, 2024. [Online]. Available: http://arxiv.org/abs/2309.15025
[153] H. Li, J. T. Moon, S. Purkayastha, L. A. Celi, H. Trivedi, and J. W. Gichoya, “Ethics of large language models in medicine and medical research,” *Lancet Digit. Health*, vol. 5, no. 6, pp. e333–e335, Jun. 2023, doi: 10.1016/S2589-7500(23)00083-3.
[154] K. Singhal *et al.*, “Large language models encode clinical knowledge,” *Nature*, vol. 620, no. 7972, pp. 172–180, Aug. 2023, doi: 10.1038/s41586-023-06291-2.
[155] “AI ushers in next-gen prior authorization in healthcare | McKinsey | McKinsey.” Accessed: Jun. 25, 2024. [Online]. Available: https://www.mckinsey.com/industries/healthcare/our-insights/ai-ushers-in-next-gen-prior-authorization-in-healthcare
[156] C. A. Grimm, “High Rates of Prior Authorization Denials by Some Plans and Limited State Oversight Raise Concerns About Access to Care in Medicaid Managed Care”.
[157] L. A. Lenert, S. Lane, and R. Wehbe, “Could an artificial intelligence approach to prior authorization be more human?,” *J. Am. Med. Inform. Assoc.*, vol. 30, no. 5, pp. 989–994, May 2023, doi: 10.1093/jamia/ocad016.
[158] “Streamlining and Reimagining Prior Authorization Under Value-Based Contracts: A Call to Action From the Value in Healthcare Initiative’s Prior Authorization Learning Collaborative | Circulation: Cardiovascular Quality and Outcomes.” Accessed: Jun. 25, 2024. [Online]. Available: https://www.ahajournals.org/doi/10.1161/CIRCOUTCOMES.120.006564
[159] “Request for Information: Electronic Prior Authorization Standards, Implementation Specifications, and Certification Criteria,” Federal Register. Accessed: Jun. 25, 2024. [Online]. Available: https://www.federalregister.gov/documents/2022/01/24/2022-01309/request-for-information-electronic-prior-authorization-standards-implementation-specifications-and
[160] “Prior authorization practice resources,” American Medical Association. Accessed: Jun. 25, 2024. [Online]. Available: https://www.ama-assn.org/practice-management/sustainability/prior-authorization-practice-resources
[161] K. A. Tran, O. Kondrashova, A. Bradley, E. D. Williams, J. V. Pearson, and N. Waddell, “Deep learning in cancer diagnosis, prognosis and treatment selection,” *Genome Med.*, vol. 13, no. 1, p. 152, Sep. 2021, doi: 10.1186/s13073-021-00968-x.
[162] A. M. Tsimberidou, M. Kahle, H. H. Vo, M. A. Baysal, A. Johnson, and F. Meric-Bernstam, “Molecular tumour boards - current and future considerations for precision oncology,” *Nat. Rev. Clin. Oncol.*, vol. 20, no. 12, pp. 843–863, Dec. 2023, doi: 10.1038/s41571-023-00824-4.
[163] A. Zehir *et al.*, “Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients,” *Nat. Med.*, vol. 23, no. 6, pp. 703–713, Jun. 2017, doi: 10.1038/nm.4333.
[164] Gomes Bruna and Ashley Euan A., “Artificial Intelligence in Molecular Medicine,” *N. Engl. J. Med.*, vol. 388, no. 26, pp. 2456–2465, Jun. 2023, doi: 10.1056/NEJMra2204787.

[165] E. R. Malone, M. Oliva, P. J. B. Sabatini, T. L. Stockley, and L. L. Siu, “Molecular profiling for precision cancer therapies,” *Genome Med.*, vol. 12, no. 1, p. 8, Jan. 2020, doi: 10.1186/s13073-019-0703-1.

[166] “A Seismic Shift: Expanding the Reach of Precision Cancer Medicine | Duke Cancer Institute.” Accessed: Jun. 24, 2024. [Online]. Available: https://www.dukecancerinstitute.org/blogs/seismic-shift-expanding-reach-precision-cancer-medicine

Glossary

For definitions of CHAI's core principles, see Section 5.

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.

Algorithm: A set of computational rules or a process to be followed in order to solve a problem.

Artificial Intelligence: A branch of computer science focused on developing techniques that enable computers to mimic intelligent behavior, akin to that of humans. The term also applies to machine-based systems that can make predictions, recommendations, or decisions, thereby influencing real or virtual environments.

Business Owner: Typically a member of the implementer team, this individual is responsible for articulating the need for a given AI solution. In certain cases, the business owner helps with the AI solution's development, tests it for performance and utility, and assesses its impact. As champion for its adoption, the business owner is key in driving the success of the AI solution.

Calibration: In the context of clinical decision support (CDS), calibration refers to the accuracy with which the predicted probabilities of outcomes match the actual observed outcomes. A well-calibrated CDS tool will provide probability estimates that accurately reflect real-world results, ensuring that the predictions are reliable and can be trusted in clinical decision-making. Calibration thus helps to ensure that risk predictions are neither overestimated or underestimated, thereby enhancing the tool's effectiveness and trustworthiness.

Change Management: A structured approach to transitioning individuals, teams, and organizations from a current state to a desired future state. In a healthcare setting, this involves activities such as planning, implementing, and monitoring changes to processes, technologies, and policies to improve patient care, enhance operational efficiency, and ensure stakeholder alignment and engagement.

Clinical Decision Support (CDS): Software that provides clinicians, staff, patients, or other individuals with knowledge and person-specific information, intelligently filtered or presented at appropriate times, to enhance health and healthcare.

Cost-Benefit Analysis: A process used to measure or calculate the benefits of a decision minus its potential financial cost. As distinct from a risk-benefit analysis, a cost-benefit analysis focuses on monetary gains and losses, usually to an organization.

Deep Learning (DL): A subset of machine learning (ML) that focuses on neural networks with multiple layers of neural networks. Deep learning employs statistics to spot underlying trends in data patterns, and it often involves training the network on vast datasets to make predictions. It requires extensive computing power and labeled data, making it prone to bias and security risks.

Developer Team: In the context of CHAI's Responsible AI Guide, this refers to stakeholders primarily involved in the AI solution development process and the maintenance of the solution.

End User: The person who actually uses and interfaces with the AI solution. In the context of health AI, that individual is likely to be a clinician, an administrator, or ops personnel, but in some cases may also be a patient.

End-of-Life (EOL): The point in time when hardware or software systems reach the end of their useful life.

Executive Sponsor: In the context of CHAI's Responsible AI Guide, this is an individual from the implementer organization's leadership, aligning the implementer team and the AI solution with the organization's strategic priorities and resources.

Generative AI: A subset of artificial intelligence that generates new content like text or images. In a healthcare setting, it may be paired with other AI models to perform tasks like history summarization or inbox responses.

Health AI: The application of algorithmic systems to a suite of tasks in the healthcare ecosystem, including decision support, diagnosis, treatment planning, medical imaging analysis, patient monitoring, clinical note taking, precision medicine, and various administrative processes.

Implementer team: In the context of CHAI's Responsible AI Guide, this is the group of stakeholders involved in implementing, using and integrating an AI solution in health system workflows.

Learning Health System (LHS): A health system in which knowledge gathering and generation is embedded in daily practice to improve individual and population health.

Machine Learning (ML): A subset of AI that focuses on models that learn from and make predictions based on data, without being explicitly programmed. Common ML techniques include supervised learning (where models learn from labeled data), unsupervised learning (where models find patterns in unlabeled data), and reinforcement learning (where models learn to make decisions by interacting with an environment or human trainers).

Natural Language Processing (NLP): A subset of AI that focuses on the interaction between computers and the comprehension of human language. NLP teaches computers to interpret and generate language for tasks like translation, text summarization, and speech recognition.

Privacy-Enhancing Technologies (PET): Tools, techniques, or methods designed to protect and enhance the privacy of personal and organizational data. PETs can include encryption, anonymization, access controls, and other mechanisms.

Probability: In the context of clinical decision support (CDS), probability refers to the statistical likelihood that a specific clinical outcome or event will occur, based on the analysis of relevant data and predictive models. Probabilities help clinicians assess potential risks and make informed decisions by quantifying the uncertainty associated with different diagnostic, prognostic, or treatment options.

Responsible AI: A growing field considering such important factors as safety, reliability, fairness, and the ethical implications of AI systems and their uses.

Risk-Benefit Analysis: In the context of health AI, a risk-benefit analysis involves evaluating the potential risks and benefits associated with the development, deployment, and use of AI solutions in healthcare settings. Unlike a cost-benefit analysis, which primarily focuses on monetary costs and benefits, a risk-benefit analysis may focus on patient safety, clinical efficacy, data privacy, ethical considerations, regulatory compliance, societal impact, and other relevant aspects of risk and benefit.

Technology Owner: In the context of CHAI's Responsible AI Guide, the technology owner on the implementer team is responsible for the technical aspects of the AI solution, including its functions and maintenance during deployment.

Trustworthiness: In the context of health AI, trustworthiness refers to the extent to which AI systems are perceived as reliable, transparent, and ethical by stakeholders including patients, clinicians, and healthcare administrators. Trustworthiness ecompasses factors such as data integrity, model robustness, accountability, and the ability to explain and justify AI-driven decisions.

Use Case: In the context of health AI, this refers to a specific application or scenario in which an AI solution is deployed to address a particular problem or achieve a defined objective.

Appendices

Appendix 1: Use Case Profiles

Predictive EHR Risk Use Case: Pediatric Asthma Exacerbation Risk

Artificial Intelligence (AI) tools intend to improve asthma control and reduce asthma exacerbations (AE) for pediatric asthma patients. The solution presents relevant asthma care information from the EHR to pediatric clinicians before a patient visit resulting in reduced EHR review time [145], [146], [147].

AI Algorithm Type

Natural Language Processing (NLP) systems extract relevant clinical concepts from notes in a patient's medical record. The AE prediction within 12 months algorithm uses a logistic regression model.

Description

Achieving optimal care for pediatric asthma patients depends on clinicians efficiently accessing pertinent patient information. Still, relevant information is often scattered throughout the patient chart in the EHR in both structured data and unstructured clinical notes. To support pediatric clinicians to automate and optimize the Asthma Action Plan (AAP) guidelines, an ML-based clinical decision support tool was developed to extract and synthesize pertinent patient data related to asthma management from the EHR and provide an AE risk score prediction. The outputs are included in an interface generated within an Asthma Exacerbation application, accessed through a patient record in the EHR, and include a 1) summary of relevant clinical information for asthma management, 2) prediction of future risk of AE (with contextualization by including relevant clinical information pertaining to a patient's asthma status), and 3) list of actionable intervention options.

End Users and Stakeholders

Pediatric clinicians such as asthma specialists, allergists, pulmonologists, and pediatricians are the main users of this application. In addition to pediatric clinicians, other stakeholders include:

- Nurse practitioners
- Nurses
- Asthma Care Coordinators/Asthma Managers
- Pediatric Patients (with a diagnosis of asthma)
- Pediatric Patient Caregivers
- EHR Vendors
- Privacy
- Security

Model Output / Decisions and Actions Made

See Figure 1. The NLP algorithm runs nightly based on a fixed cohort of patients with an asthma diagnosis reported in the EHR. The NLP system extracts ~20 relevant clinical concepts from notes (structured and unstructured data). The logistic regression model for AE risk runs when a patient chart is launched in the EHR and outputs a value between 0 and 1. A cutoff was established via discussions with clinicians, so the output value is shown as either high or low risk rather than a value between 0 and 1 (the numeric value is not shown on the Asthma Exacerbation

application interface). The NLP algorithm and logistic regression model outputs are input into an interface generated on the Asthma Exacerbation application, that sits outside the EHR. A pediatric clinician can access the application through a patient's chart in the EHR to review the model outputs on the Asthma Exacerbation application interface before a patient appointment.

A standardized list of actionable intervention options for a pediatric clinician to review is also presented on the Asthma Exacerbation application interface. Relevant patient data for each intervention option is populated from the patient's medical record, and a pediatric clinician can click on possible intervention options to review patient data associated with each. The pediatric clinician will decide on treatment options (asthma care plan) leveraging their review of the interface content. The available decisions are: 1. No change in asthma care, 2. Refer to Pediatric Asthma Coordination Program (AMP) coordinator, 3. Address risk factors through the Pediatric Asthma Coordination Program (AMP), 4. Refer to community health worker, 5. Recommend Allergy test, 6. Recommend Spirometry, 7. Recommend asthma-specific regular visits (e.g., three months), 8. Complete the Asthma Action Plan (AAP), 9 Complete the Asthma Control Action Plan (ACA), 10. Complete Asthma Control Test (ACT), 11. Medication step-up, step-down, or maintain, and 12. Other recommendations.

Interface, Application, & Technological Environment

The asthma care plan automation environment relies on the aggregation of patient data from the EHR. The summary of relevant information and AE risk score are securely populated into an interface on the Asthma Exacerbation application accessible to pediatric clinicians. The application also contains options to open new panes, which provide detailed information such as asthma status, risk factors, asthma exacerbation risk, and factors contributing to high risk. The computed AE risk score predictions are stored in a database. The most recent previous stored prediction is returned if nothing new is documented in the patient record (i.e., patient encounter).

Privacy & Security

The healthcare organization's secure instance of EHR deployment to access and authenticate the patient record is leveraged for the product's environment. The patient data are stored in the EHR, and a pediatric clinician's access to a patient record is secured and monitored. The Asthma Exacerbation application can only be accessed by launching from the EHR. Access to the data collected is limited to only those personnel who need access, and proper handling of data is defined per the healthcare organization's policy. Software development included software configuration management and instructions on how to best install and configure the product were documented. Software testing for all components of the product was conducted, including boundary conditions to mitigate the impact of software errors.

Data sources and training

The logistic regression model for AE risk score prediction was trained using a study cohort of 1,889 patients from a pediatric asthma registry at a healthcare organization as of 2021-06-08. A patient's clinical notes were extracted between 2018-05-05 and 2021-06-08. All the clinical note dates of a patient are considered as visit dates. For modeling, each patient's data is partitioned

into two categories 1) Prediction window 2) Observation window. A decision date is a time point where the prediction window and observation window are separated. For a 12-month prediction window, we calculate the decision date by taking the latest visit date for the patient and moving back 12 months to mark the decision date. Any data before the decision date is considered an observation window.

Data Flow (Input/Output) / Pre-conditions

Preconditions include that a patient must have a diagnosis of asthma in the EHR, patient data must be available in the EHR, and pediatric clinician chart review and treatment decision making can be done leveraging the summary of clinical information, AE risk score, and list of actionable intervention options.

Basic flow

The AI Tool for Pediatric Asthma Exacerbations is intended to be used on patients with a documented diagnosis of asthma. The tool maintains a list of patient MRNs and for each of those MRNs, it leverages patient data documented in the EHR to fetch NLP data, compute AE risk prediction, and store data. The NLP algorithm is run every morning (generating enriched patient notes) and the logistic regression prediction model is run when the patient chart is launched in the EHR.

An appointment is scheduled by a patient with an asthma diagnosis. Prior to a pediatric clinician having the appointment with an asthma patient, the clinician launches the asthma patient record in the EHR. The clinician then clicks the Asthma Exacerbation application in the patient record and an interface is generated with the AI outputs, which includes a summary of relevant patient information, an AE risk score (high or low), and list of actionable intervention options. On the interface, the clinician can review the relevant clinical information pertaining to a patient's asthma status, including those contributing to the AE risk score and the patient data that may contribute to possible intervention options. Prior to the patient visit, a pediatric clinician will review the interface generated by the Asthma Exacerbation application. The pediatric clinician and patient (and caregiver) will discuss the asthma care plan.

Alternative flow

Pediatric clinicians manually review patient data in the EHR and will make decisions on asthma care and treatment options without the AI tool.

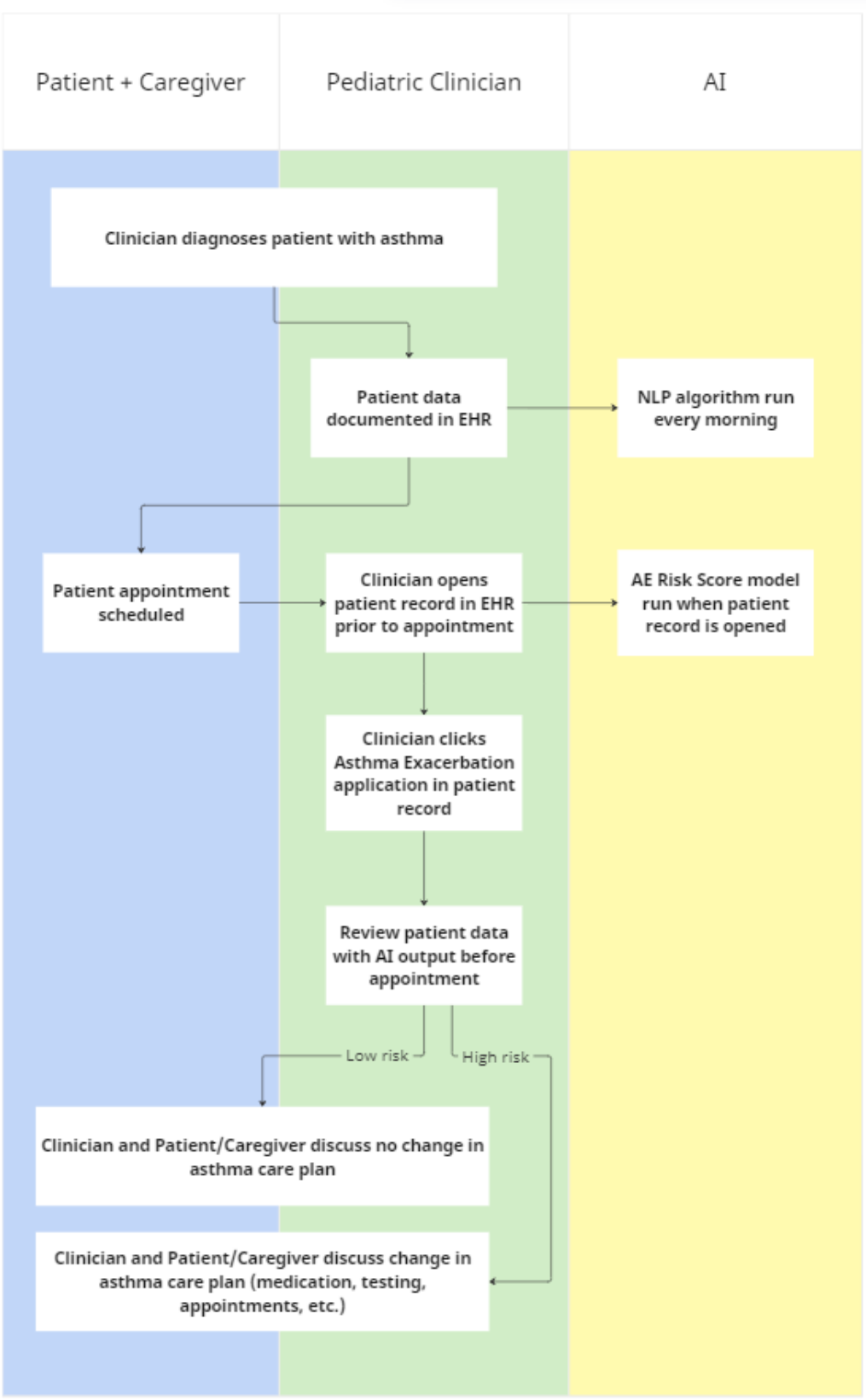
Limitations

The AI Tool for Pediatric Asthma Exacerbations is intended to automatically extract and synthesize large amounts of pertinent patient data related to asthma management from EHRs to support pediatric clinicians to optimize asthma care through the following: 1) summary of the most relevant clinical information for asthma management (NLP system), 2) prediction of future asthma exacerbation risk using a logistic regression prediction model, and 3) list of actionable interventions. The Asthma Exacerbation application interface outputs are intended to supplement

a pediatric clinician's evaluation and care plan rather than take precedence, which must be documented through labeling and in training materials provided to pediatric clinicians. The AI Tool for Pediatric Asthma Exacerbations is intended to be used only on pediatric patients between 6-17 years old with a diagnosis of asthma in the EHR who receive pediatric primary care at a healthcare organization. Risks of the AI Tool for Pediatric Asthma Exacerbations include:

- False positive of the logistic regression model (AE risk score), meaning that AI miscalculates/misclassifies a patient's AE risk score, classifying the patient as high risk but the patient is actually low risk. The incorrect false positive could be read by a pediatric clinician and leads to mismanagement of patient asthma, which can impact the asthma care plan (i.e., reevaluating medication, scheduling testing/treatment interventions) when not necessary.
- False negative of the logistic regression model (AE risk score), meaning that AI miscalculates/misclassifies a patient's AE risk score, classifying the patient as low risk but the patient is actually high risk. The incorrect false negative could be read by a pediatric clinician and leads and leads to mismanagement of patient asthma, which can impact the asthma care plan (i.e., patient not receiving time-sensitive treatment, reevaluating medication) if necessary action is not taken.
- Incomplete, ambiguous, or no result of the logistic regression model (AE risk score), resulting in a delay in AE risk score output, which then increases the wait time for clinician review and/or communication with the patient, and an error value indication ("unavailable") is shown.
- Pediatric clinician misinterpretation of logistic regression model (AE risk score) output, indicating there is an unknown meaning of what the thresholds mean and what constitutes a high and low AE risk score classification for a patient.
- NLP algorithm generating enriched patient notes misses relevant patient data or hallucinates patient data, resulting in incorrect patient related information (asthma symptoms, asthma control status, lung function tests) displayed to a pediatric clinician on the Asthma Exacerbation application.
- Breach of confidentiality of patient health information due to unauthorized access of the patient record, missing data, and bias in training dataset (patient demographic tool is being used on is not included in training dataset).
- Pediatric clinician overreliance on AE risk score, leading to automation bias.

Figure 1: Asthma Exacerbation Risk Use Case Swimlane Diagram



Imaging Diagnostic Use Case: Mammography

Use Case and Goals

Several AI-based algorithms have been developed to assist radiologists who interpret screening mammograms looking for breast cancer with an AI-augmented "second pair of eyes". These algorithms perform image analysis and have been shown to detect abnormalities on mammograms with similar accuracy as radiologists and then triage or prioritize abnormal mammograms over normal mammograms for more timely interpretation. These tools have the potential to improve the breast cancer screening workflow - patients and their families may benefit from quicker results. In addition, radiologists as the primary end users may benefit from increased decision-making confidence while interpreting screening mammograms alongside AI tools and improved workflow efficiencies [125], [148], [149], [150].

End Users and Stakeholders

As with most clinical workflows, there are many collaborators needed for processing mammography screenings with a health system.

- Screening mammography patients and their family members.
- Clinicians who refer women for screening mammograms (primary care).
- Physicians who treat women with breast cancer (oncologists, surgical oncologists, and plastic surgeons).
- Radiologists, Radiology technologist, and radiology administrators.
- Healthcare risk management.
- Healthcare privacy and security and IT teams.
- Radiology mammography accreditation and regulatory bodies (ACR,FDA, MQSA).
- Groups who publish breast cancer screening guidelines (USPSTF, ACS, ACR).
- Health AI vendors.

In this example, the AI algorithm is third-party, vendor-developed and has been cleared by the FDA. The AI system uses deep learning technology in the form of a convoluted neural network to identify regions on screening mammogram images that are suspicious for cancer. Most women who participate in screening mammography do not have breast cancer, so prioritizing the subset of patients who have an abnormal mammogram is valuable for triaging these patients for further workup. This AI system assigns each mammogram a malignancy risk score on a continuous scale ranging from 1 to 10. Cancer prevalence increases sharply in the risk score 10 group, allowing mammograms with potentially suspicious findings to be prioritized for interpretation. The risk score enables the prioritization of abnormal mammograms in the screening workflow.

Description

The AI tool identifies and highlights regions on mammogram medical images that are suspicious for cancer and assigns a malignancy risk score to the finding.

Model Output / Decisions and Actions Made

Radiologists interpret screening mammogram images alongside AI output in the form of image annotation and workflow prioritization (see figure 1). The process annotates overlay/computer-aided "detection marks" on the mammogram image that highlight potential abnormalities and displays an exam-specific malignancy "risk score" in the form of a numerical score or classification of low, medium, and high risk. For example, mammograms with higher risk scores can be prioritized for interpretation ahead of mammograms with lower risk scores. However, all mammograms are ultimately interpreted by the radiologist regardless of risk score. The radiologist communicates their findings by generating a standardized report according to the Breast Imaging Reporting and Data System (BI-RADS) developed by the American College of Radiology (ACR) that includes the assignment of a BI-RADS score and recommendations for follow-up imaging and/or screening frequency as applicable. This report is released to the patient and ordering clinician upon signing. Decisions and actions made by the radiologist during the screening mammography workflow including the interpretation of screening mammogram images and generation of the radiology report are performed by the radiologist and not observed by patients.

Interface, Application, & Technological Environment

Digital screening mammography (low energy x-ray evaluation of breast tissue) records an image of the patient. The system securely transmits the image to a medical image viewing platform (PACS - picture archive and communication system) and then to the vendor's on-site server for PHI removal. The server sends the de-identified images through the provider's perimeter firewall to the vendor's cloud-based platform where the AI tool analyzes the images, generates annotation marks, and calculates prioritization scores. The platform transmits the information back to the vendor's on-site server where software links the PHI, results, and image. The radiologist views the output on the PACS or the vendor-specific image viewer. This process typically takes several minutes to complete.

Privacy and Security Information

- Access controls for devices to mitigate compromise of diagnostic system components and possible direct evasion attacks which could add noise to images.
- Access controls and encrypted API transmission channels to cloud processing and storage.
- Monitor privacy and security contractual requirements with third parties.
- Lack of data integrity could lead to inaccurate prediction and possible patient harm.
- Access controls and federated learning to protect against data disclosure.
- Enlarge training and pre-processing cleaning of datasets to protect against data poisoning.
- Use reliable sources, such as qualified radiologists, to label training data.

Data sources and training

The algorithm was trained and tested using more than 200,000 mammograms acquired at multiple institutions representing more than ten countries, a range of populations, modality manufacturers, and variations in workflows. More than 10,000 biopsy-proven cancers are annotated in this dataset. (Lang et al, Lancet Oncol 2023)

Data Flow (Input/Output) / Pre-conditions

The source mammogram image data must flow automatically from equipment source to PACS to the vendor and back to PACS. Radiologists must be logged into PACS and a dictation software system.

Basic flow (see Figure 1)

Screening mammogram images are obtained at source and images are sent electronically to PACS, images are sent from PACS to health AI vendor for analysis and annotation, annotated images and prioritization flag sent back to PACS for viewing, radiologist must be logged into PACS to view prioritization flag, images, annotated images and dictate their interpretation.

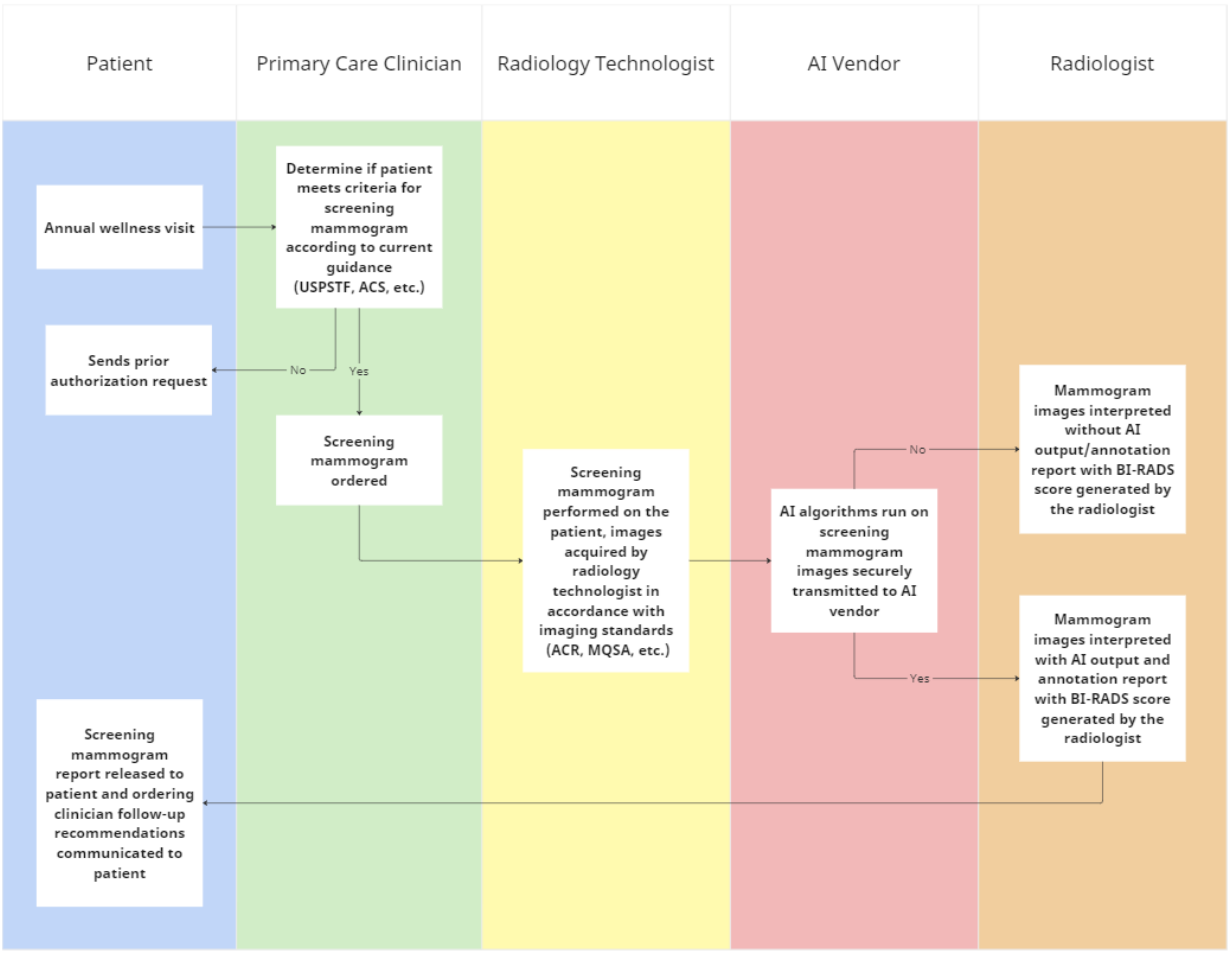
Alternative flow

Radiologist interprets mammogram images without AI tool.

Limitations

Compatibility with local systems such as PACS (there are a wide variety of PACS systems in clinical use) can interfere or limit functionality of the algorithm. Algorithm performance can differ in local populations that may vary in composition from the training populations. For example, algorithms are typically trained on highly selected “cancer-enriched” datasets that are not necessarily representative of the general screening population. Anchoring bias is a risk for this type of AI solution, which occurs due to over-reliance on what the AI identifies, leading to less attention being given to areas not flagged by the system.

Figure 2: Imaging Use Case Swimlane Diagram



Generative AI Use Case: EHR Query and Extraction

Paradigm Use Case and Goals

Natural Language Querying of Electronic Medical Records (EHRs) enhances the process of extracting information from lengthy, noisy, and unstructured information from these systems to facilitate clinical decision-making [151], [152], [153], [154].

AI Algorithm Type

The query and extraction process utilizes Large Language Models (LLMs), including decoder-only models like GPT, encoder-only models such as BERT, encoder-decoder models like T5, or an ensemble of models. These models may be designed as generalists or specifically tailored or fine-tuned for biomedical content.

End Users and Stakeholders

These AI tools aid healthcare end users such as clinicians respond to patient messages in the EHR. Stakeholders affected by this tool include:

- Healthcare providers,
- Patients,
- Healthcare administrators,
- Health AI vendors,
- Large Language Models (LLM) and Generative AI vendors, and
- EHR vendors.

Description

This process allows healthcare professionals to query a patient's complete clinical record, including unstructured notes and other EHR data, using a natural language interface. For example, during or after a clinical encounter they can ask the system "since when has this patient been on beta blockers?", "does she have any family history of cancer?", or "what's his more recent h1c?". This enables efficient and intuitive information extraction that saves time for clinicians and patients.

Model Output / Decisions and Actions Taken

The system generates a response in the form of human-readable natural language text, comprising a sequence of words that answer the user's query. It also provides specific references to sources, including links to actual clinical notes or data elements that support the answer. For example, when asked since when a patient has been on beta blockers, the system responds in natural language (i.e. "He was first prescribed Carvedilol 20mg daily on May 5th, 2018") and provides a hyperlink to the specific prescription record (in case the doctor wants to verify this or read the visit summary from that date for more background).

Interface, Application, & Technological Environment

The user interface can take two forms: a web interface or a widget integrated within the existing EHR software. Users input queries via keyboard or utilize speech recognition (speech-to-text), with the system returning responses either as on-screen text or by speaking in a synthetic voice (text-to-speech).

Privacy and security

The primary function of the AI Large Language Model is to assist healthcare professionals in extracting patient information dispersed within the electronic health record. Its role is not to provide clinical recommendations, whether diagnostic or therapeutic. This approach mitigates risk, as clinicians can verify the sources (context) upon which the answers to their queries are based.

The main risk involved is the occurrence of false negatives (low recall). For instance, a nurse might consult the Large Language Model about a patient's allergies before administering an aminoglycoside (an antibiotic). If the information is recorded in the clinical record but the LLM fails to retrieve it, the nurse might overly rely on the LLM's response and administer a drug to which the patient is allergic.

Possible negative impacts of hallucinations, such as false positives where the LLM suggests a patient has a condition not supported by the electronic health record, are effectively mitigated. This is because the LLM's responses are always accompanied by pieces of contextual information from the EHR, which form the basis of the answer. This provision of contextual data alongside the LLM's response serves to mitigate the risk derived from eventual false positives thanks to human supervision.

Regarding privacy concerns, since the model accesses sensitive patient information from electronic health records, it is imperative to ensure that all data handling complies with healthcare privacy regulations such as HIPAA in the United States or GDPR in the European Union.

To ensure patient privacy, the LLM must incorporate robust security measures that thwart unauthorized access to patient data. These measures include employing secure data transmission protocols, encrypting data both at rest and during transit, and implementing stringent access controls to guarantee that only authorized healthcare professionals can interact with the system. Additionally, the Safeguard LLM module plays a crucial role in scrutinizing user queries. It evaluates whether a query is reasonable, relevant, and contextually coherent within the framework of a clinical encounter. The module may also consider the professional role or specialty of the user submitting the query. For example, it might refuse to provide information about sexual paraphilias to individuals in unrelated fields, such as radiology technicians or ophthalmologists. This tailored approach to query management further strengthens the safeguarding of sensitive patient information.

The fundamental design philosophy behind the system is to serve as a universal repository of valuable information accessible to various stakeholders. However, not all information recorded in the EHR is relevant or suitable for every clinical context. Therefore, there is a need for a mechanism, such as the Safeguard LLM, that acts as an intermediary. This mechanism helps to

align user requests to the Large Language Model connected with the EHR, ensuring compliance with the organization's defined policies or rules on information access.

Data sources and training

Large language models are typically trained in a self-supervised manner, drawing from vast content sources, such as publicly available internet data. In some instances, they are specifically trained to encompass biomedical content, or may also include private data, such as private clinical guidelines or patient clinical records. Multimodal models may integrate various data types, including images, audio, or domain-specific sequential content like physiological signals or genomic and proteomic information.

Data Flow (Input/Output) / Pre-conditions

The system requires real-time access to all pertinent patient information within the EHR. User authentication and legal authorization to access patient data are mandatory prerequisites. All interactions with the system, including queries and LLM responses, must be logged and treated as private medical content access.

Basic flow (see Figure 1)

Backend Data preprocessing and information Extraction:
Initially, and whenever new data is added to the Electronic Health Record (EHR), this information is transmitted to the Large Language Model (LLM) Encoder. The Encoder then creates an embedded representation of this data, which is subsequently stored in a specialized database, such as a vector database or a knowledge graph.

Clinical encounter:

The process begins when a patient consults a healthcare professional, marking the start of a clinical consultation. In this interaction, the healthcare professional gathers information to form an initial understanding of the patient's condition. Should the clinician seek support from the Artificial Intelligence (AI) System, they can make inquiries in natural language about information stored in the patient's Electronic Health Record (EHR). These inquiries are then submitted to the Safeguard LLM. It's important to note that the Safeguard LLM is programmed to reject any prompts requesting medical recommendations, such as differential diagnoses or treatment and management plans, as these requests fall outside the system's intended scope. The Safeguard LLM's primary role is to ensure that each query is appropriate and aligns with the expected use of the system. This Safeguard LLM is also responsible for ensuring that the information requested about a patient is appropriate for the context of the clinical encounter and the healthcare professional's profile. For example, it could be utilized to decline inquiries about sexual preferences or religious beliefs when an ophthalmologist is using the system in the context of cataract surgery.

Artificial Intelligence Decision Support:

Should the Safeguard LLM classify a query as inappropriate — particularly those seeking medical recommendations like differential diagnoses or treatment plans, which are beyond the system's capabilities — it will issue a rejection notice. In these instances, the clinician will rely exclusively on their own expertise, marking the end of the AI-assisted process.
If, however, the query is suitable, primarily focusing on information from the Electronic Health Record (EHR), it progresses to an LLM sub model named the "Encoder." This sub model transforms the query into an embedded representation. These embeddings are then analyzed by the RAG Module, another LLM submodule, which is tasked with retrieving relevant EHR data to address the query[1].

Subsequently, the raw EHR data, constituting the context, is compiled, and forwarded to a different LLM submodule called "Decoder," along with the refined query. The Decoder formulates a response, adhering strictly to the information available in the EHR, and sends this back to the Safeguard LLM for a final appropriateness check[2,3]. If the response is found to be lacking, a rejection message is sent to the clinician, who then concludes the process based on their clinical judgment.

Conversely, if the response is appropriate, it is delivered in natural language[4], supplemented with references to the EHR data used in the response. The healthcare professional then reviews this information, verifies the sources if necessary, and integrates these insights into their clinical decision-making process, while informing the patient about the role of the AI system in their clinical evaluation.

Alternative flow

Healthcare professional searches for the answer to the question in the electronic health record by manual review, or using simpler search tools, before taking the clinical decision or making the decision without having that information.

Limitations

Integration:
Integrating LLMs with EHRs can be challenging, and the process may not be easily replicable across different EHR systems. The “EHR" itself may not be one system: it may be necessary to retrieve data from separate inpatient EHR, outpatient EHR, laboratory, pharmacy, and billing system to create a unified and complete view of a patient.

Algorithm Generalizability & Accuracy:
The model's performance can vary depending on the language used to document patient encounters and, even within the same language and care setting, due to a specific jargon of an organization or specialty. This variability makes it more challenging to generalize and validate the safety and accuracy of the system’s responses in different contexts.

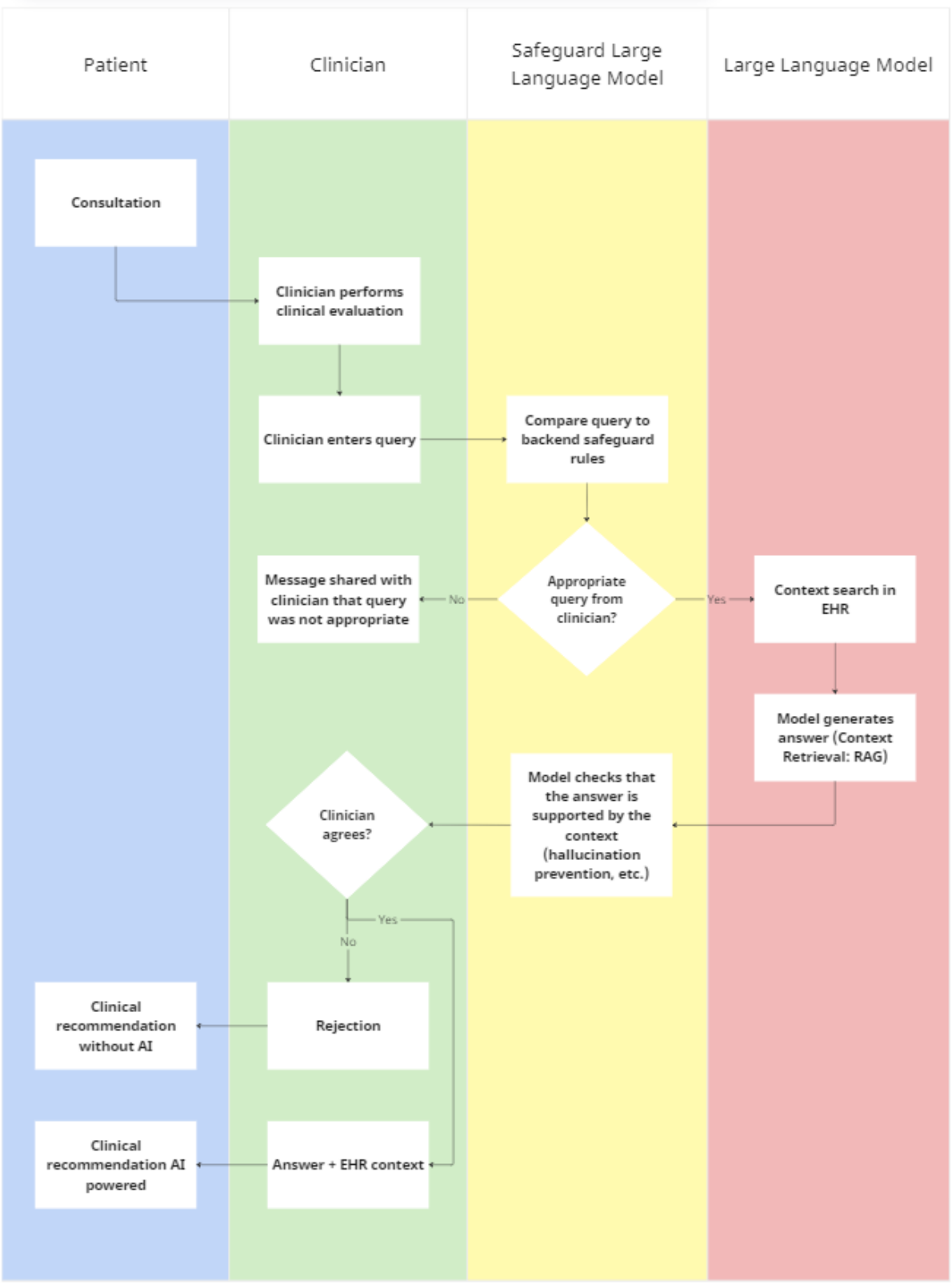
The main risk involved is the occurrence of false positives (low recall). For instance, a nurse might consult the Large Language Model about a patient's allergies before administering an aminoglycoside (an antibiotic). If the information is recorded in the clinical record but the LLM fails to retrieve it, the nurse might overly rely on the LLM's response and administer a drug to which the patient is allergic.

The risk of hallucinations, such as false positives where the LLM suggests a patient has a condition not supported by the electronic health record, is relatively lower. This is because the LLM's responses are always accompanied by pieces of contextual information from the EHR, which form the basis of the answer. This provision of contextual data alongside the LLM's response serves to mitigate the risk of errors arising from false positives.

Information Extraction vs. Clinical Recommendations:
While this use case focuses on retrieving information that exists in the EHR in a faster and more intuitive way, a natural extension of the same user interface – with higher risk – is also to provide clinical guidance. For example, this would extend to answering questions like “what would you prescribe this patient?”, “which lab orders would you recommend?”, or “which trials is this patient a candidate for?”. However the primary function of the AI Large Language Model is to assist healthcare professionals in extracting patient information dispersed within the electronic health record. Its role is not to provide clinical recommendations, whether diagnostic or therapeutic. This approach mitigates risk, as clinicians are able to verify the sources (context) upon which the answers to their queries are based.

Computational Requirements:
Current state-of-the-art LLMs a far more compute-intensive compared to typical clinical information systems or EHRs. This introduces additional costs, complexity, IT overhead, and a higher risk of Personal Health Information (PHI) breaches.

Figure 3: Generative AI Use Case Swimlane Diagram



Claims-Based Outpatient Use Case: Care Management

Use Case and Goals

The claims-based Comprehensive Care Management Model seeks to improve overall, or disease-specific, care management by identifying individuals at highest risk of morbidity and health system utilization as proxies for care coordination needs.AI Algorithm Type
Predictive and/or classification-based machine learning model.

Description

This model uses local data on basic patient demographics (gender, age, race/ethnicity, geography), medical/pharmacy claims, diagnoses (as available), prior resource use, and population markers to predict:

1. current or future probability of high cost, high utilization, and risk of hospitalization,
2. future mortality, and
3. care coordination risk categories (general and clinically meaningful but not disease-specific).

The outcomes predicted by the model serve as proxies for healthcare needs more broadly.

End Users and Stakeholders

The primary users of the model are care management coordinators, case managers, clinicians (nursing, PA, physicians), community care coordinators, and executives. Groups affected by the model inside and outside the organization include:

- data analysts/scientists,
- insurance executives (quality and safety officers, population health officers, equity officers),
- patients,
- caregivers,
- privacy and security and IT teams,
- legal teams,
- hospital executives, and
- health AI vendors.

Model Output / Decisions and Actions Made

The model outputs the probability of concurrent and future risk for each predictive model (e.g. utilization, cost, hospitalization, frailty, morbidity, etc.) along with the initial risk categorizations based on predictive model outputs (rule-based). The data outputs are combined with measures of local social determinants of health using a rule-based algorithm. This process reduces the likelihood of biased estimates based on utilization-based measures. The result creates overall care coordination need and risk categories.

Data analysts from the quality team send a weekly roster of high-risk tier patients upon enrollment or re-determination to the enhanced care management team. Care coordinators use this list to prioritize outreach to newly enrolled or eligible members not currently enrolled in the

enhanced care management program to conduct further evaluation, beginning first with review of patient profile. Upon contact with members, enhanced care management staff conduct additional interviews to determine if a high-risk individual is eligible for services based on a rule-based decision tree. If eligible, a member is enrolled in enhanced care management program.

Interface, Application, & Technological Environment

The system integrates data from multiple sources, locations, and coding standards. Data are fed from primary and secondary care records into a central database. Inputs can be customized with elements such as socio-economic or functional living status. A dashboard displays population health/public health report tools and applications to payer care managers and administrators.

Privacy and Security Information

- Access controls to prevent patient health information disclosure when IT, Quality, and Care Management Teams manage data.
- Encrypted transmission channels between systems.
- Monitor third parties through audits to ensure they implement privacy and security controls required by contractual agreements.
- Weak data integrity controls could allow threat actors to manipulate data producing inaccurate risk prediction and affect care management.
- Access controls and federated learning to protect against data disclosure.
- Enlarge training and pre-processing cleaning of datasets to protect against data poisoning.

Data sources and training

Data sources include claims, encounter, pharmacy, broader SDOH, and eligibility data. Over 200 different variables are used to calculate future risk. Weights that increase or decrease future risk are based on training in over 700,000 patients.

Data Flow (Input/Output) / Pre-conditions

Patient data, medical, and pharmacy claims data must be deidentified and automatically flow from claims data storage. Future risk predictions require 1-2 years of available data.

1. Model tuned to local data.
2. Data input into models.
3. Model produces member-level output for predicted values and stratifies members into risk categories (monthly).
4. Roster of high, moderate, low risk members provided to enhanced care management team (weekly based on new enrollments)
5. Care coordinators and case managers reach out to members to conduct additional evaluation.
6. Final enhanced care eligibility decision made (member enrolled in relevant programs or not)
7. Moderate-risk members are contacted next in prioritization order.

Basic flow (see Figure 1 at the end of this document)

The patient enrolls in the health plan and completes their health assessment. Data from their enrollment and health assessment, as well as any past claims, diagnosis, etc. will be extracted by the IT team and sent to the quality team monthly. The quality team will then input this data into the predictive model (AI/ML) which will produce risk probability estimates. Based on probability cut-offs, individuals will be categorized into high, moderate, and low-risk groups. Adjustments to individual's group categorization will be made using a rule-based algorithm that adjusts for added data such as social determinants of health and other potential risk factors. Each care management program will receive a list of high and moderate risk members in accordance with model predictions and other eligibility criteria (e.g. specific diagnosis required, disability status, etc.) and will use this list to prioritize resource allocation and outreach to members for further evaluation to determine final eligibility for specialty care management program enrollment.

Alternative flow

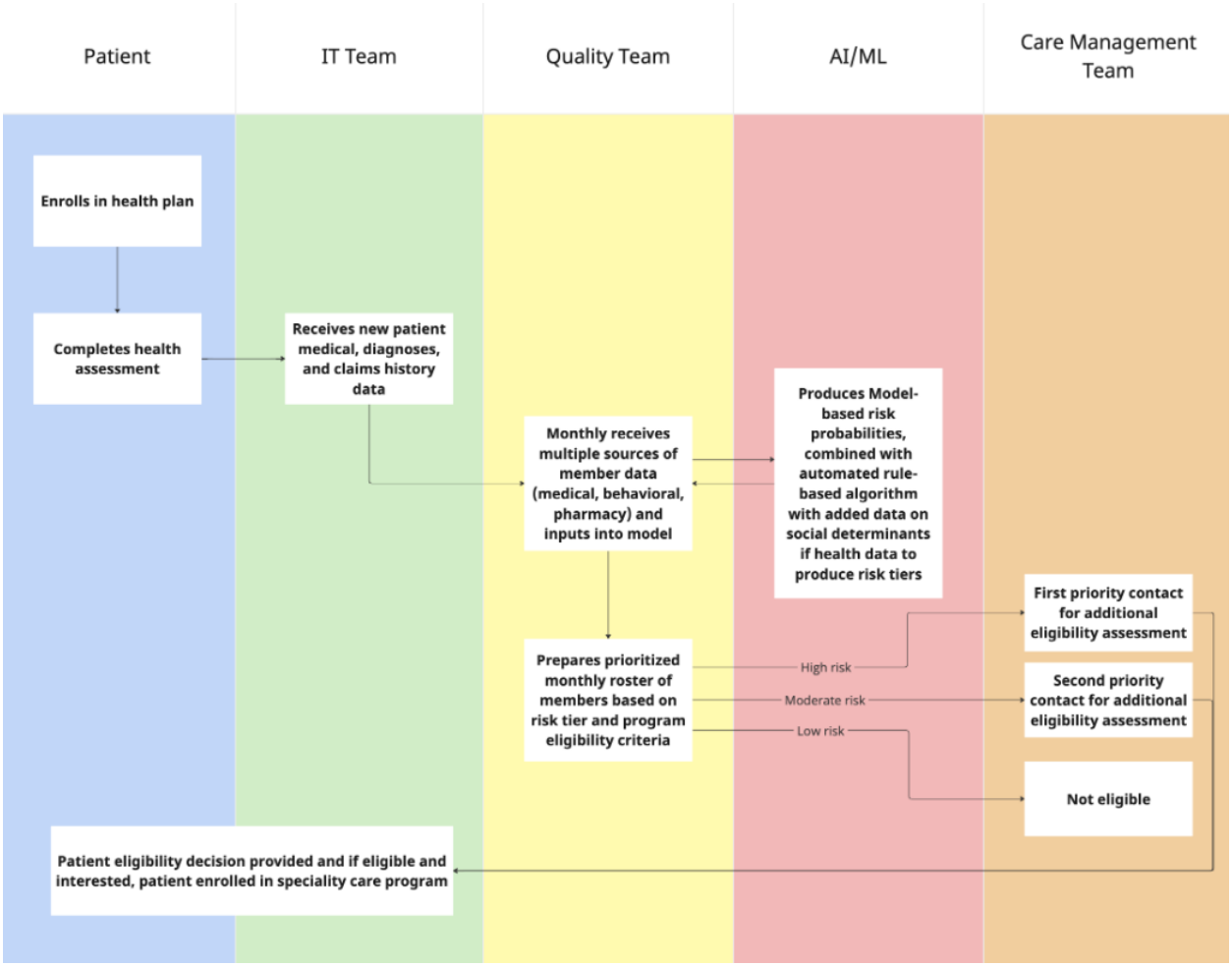
Care coordinators and case managers would either have no method for prioritizing patients or would make prioritization decisions based on simpler rule-based decisions on their own without the AI tool. This does not allow for consideration of clinically relevant co-morbidities and multi-morbidity categorization and relies primarily on specific diagnoses.
Limitations

Limitations

- Patients with chronic health conditions such as cancer may be misclassified as high risk of complex needs, for example, due to high cost and utilization but may not need enhanced care coordination.
- The model must be re-tuned to the local sample to avoid errors and inconsistencies in performance. The absence of local indices of social determinants of health in training data may lead to bias in certain racial/ethnic, or demographic subgroups.
- Missing or otherwise unavailable claims data may result in misclassification or low-risk classification. This may be more likely in marginalized socio-demographic subgroups or those living in rural areas with less access to care leading to systematic deprioritization of care resources and more barriers to care access.
- High-risk categorization applies to the top 5% of individuals with the likelihood of coordination issues due to limited resources. Other individuals at similar risk levels who fall outside that 5% may be eligible for services but not prioritized.
- Decisions made to change use context or the targeted population outside the intended or initially evaluated use can alter the effectiveness of the algorithm and require re-evaluation.
- End-user behavior and compliance can alter model effectiveness and impact members. Some notable end-user tendencies that could alter model effectiveness and performance and should be evaluated include:
    - Ambiguity/uncertainty aversion can lead to lack of use or improper use.
    - Algorithm aversion can lead to lack of use or improper use.

- Automation bias (defaulting to using AI-generated predictions or risk categories without further evaluation) may lead to inappropriate or ineffective use.
- Lack of validation around real-world clinical or operational impact may result in diminished trust, use-drift (less use or less compliant/effective use over time), and potentially wasted resources.
- Lack of clear, simple, and easily accessible (in the moment) “appropriate use instructions”, can result in cognitive short-cuts, decision fatigue, and inefficiencies that could alter effectiveness of model implementation, correct use, and ease of use.

Figure 4: Claims-Based Care Management Use Case Swimlane Diagram



Clinical Ops & Administration Use Case: Prior Auth with Medical Coding

Use Case and Goals

Implementing AI has the potential to streamline the inefficiencies of the prior authorization (PA) process in healthcare. PA seeks to align the most appropriate treatment for the patient with the best use of resources. The main stakeholders have different goals, but they all want to ensure that patients receive the right care at the right time [155], [156], [157], [158], [159], [160].

AI Algorithm Type

The PA process uses different algorithm types for the triage and the authorization automation steps. Triage determines the complexity level of the request with rules generated using classification algorithms. The authorization engine uses:

- dynamic rules for low complexity level;
- AI algorithms on claims and electronic health record for mid;
- AI and natural language processing algorithms on claims and electronic health record for high; and
- AI helps organize facts, aiding the health professional decision for very high complexity.

End Users and Stakeholders

Primary end users of the prior authorization system are clinicians, hospitals, clearing houses, and payers. Internal and external parties with an interest in the automation of PA processes:

- Clinicians
- Patients
- Hospitals
- Clearing Houses
- Payers
- Health AI vendors
- EHR vendors
- Provider-patient communication system vendors

Low complexity may use standardized technology such as FHIR for EHR data extraction and CQL for rule representation. Mid and high complexity would use deep learning (DL) neural network Transformer-based and architecture to analyze submitted information. NLP for high complexity *extracts, interprets, manipulates, and assimilates unstructured or structured spoken or written data*.

Description

Prior authorization, precertification, and prior approval refer to the same process requiring physicians and other health care providers to receive permission for reimbursement of medical service from a payer. Prior authorization, a core administrative procedure, acts as a cost control method that can prevent wasting resources on inappropriate care.

Model Output / Decisions and Actions Taken

AI models interpret authorization requests trained on prior adjudications of authorization data and the patient's medical history. Clinicians and payers review the interpretations alongside the AI output. Prior authorization can be grouped into three levels of decision-making; sometimes referred to as PA triage. Automated PA commonly uses rules-based decision-making in the first level. Rules process simple requests such as a patient's payer plan and eligibility for treatment. Artificial intelligence emerges in the second level of decision-making. Machine learning and natural language processes can automate the complex process of treatment plan request, review by payer staff, and any appeals by the clinician. The last level of decision-making includes peer-to-peer review. Proposed models show the peer-to-peer process could be automated with AI.

Interface, Application, & Technological Environment

The environment for prior authorization automation relies heavily on the aggregation of health and claims data from diverse sources. Electronic patient records from EHRs, linked prior adjudications of authorization, and personal health records among others. These applications must be able to exchange data using common application programming interfaces (APIs). Standards such as FHIR proposed in legislation or native APIs can be used to aggregate data used for training models and creating algorithms as well as in the decision-making process. Artificial intelligence in this environment requires training data, deep learning software, AI models, and natural language data extraction. This information can then be fed into rules-based and AI decision-making. Cybersecurity for prior authorization requires strong technical and administrative controls. Implementations use standard API authentication and authorization measures to limit access and scope of data to be transferred. Organizations must also screen authorized users and audit the environment for unauthorized access, breaches, and data manipulation.

Privacy and Security Information

- AI transparency in PA algorithms: Clinical Quality Language provides some transparency allowing providers to understand the basis for prior authorization decisions but it is inconsistent across all requests.
- Bias/Fairness: authorized claims may prejudice, or favoritism toward a group either intentionally or by error. In addition, there is a low representation of underserved populations in historical claims training data.
- Multiple data sources: The use of multiple data sources in PA algorithms increases the risk of data manipulation by both authorized and unauthorized users during the Engineer Data, Develop, and Deployment stages of the AI lifecycle.
- Data security/privacy – opportunities for breaches sending data between provider and payer systems.

Data sources and training

The model training process uses patient medical histories from EHRs and other sources such as HIEs linked to payer authorization claims and decisions.

Data Flow (Input/Output) / Pre-conditions

Access to patient medical data, communication with payer treatment guidelines, human review and decision-making.

Basic Flow

The provider sends treatment authorization requests from an EHR or other electronic system to the payer system. The triage engine categorizes the request according to complexity. The PA automation engine returns low, mid, and high-complexity requests back to the provider with an automated decision. This engine uses dynamic rules for low AI algorithms on claims and the patient's electronic record, and AI and NLP algorithms on claims and the patient's electronic record. AI organizes facts on very high complex requests and sends them back to the provider for manual review.

Alternative Flow

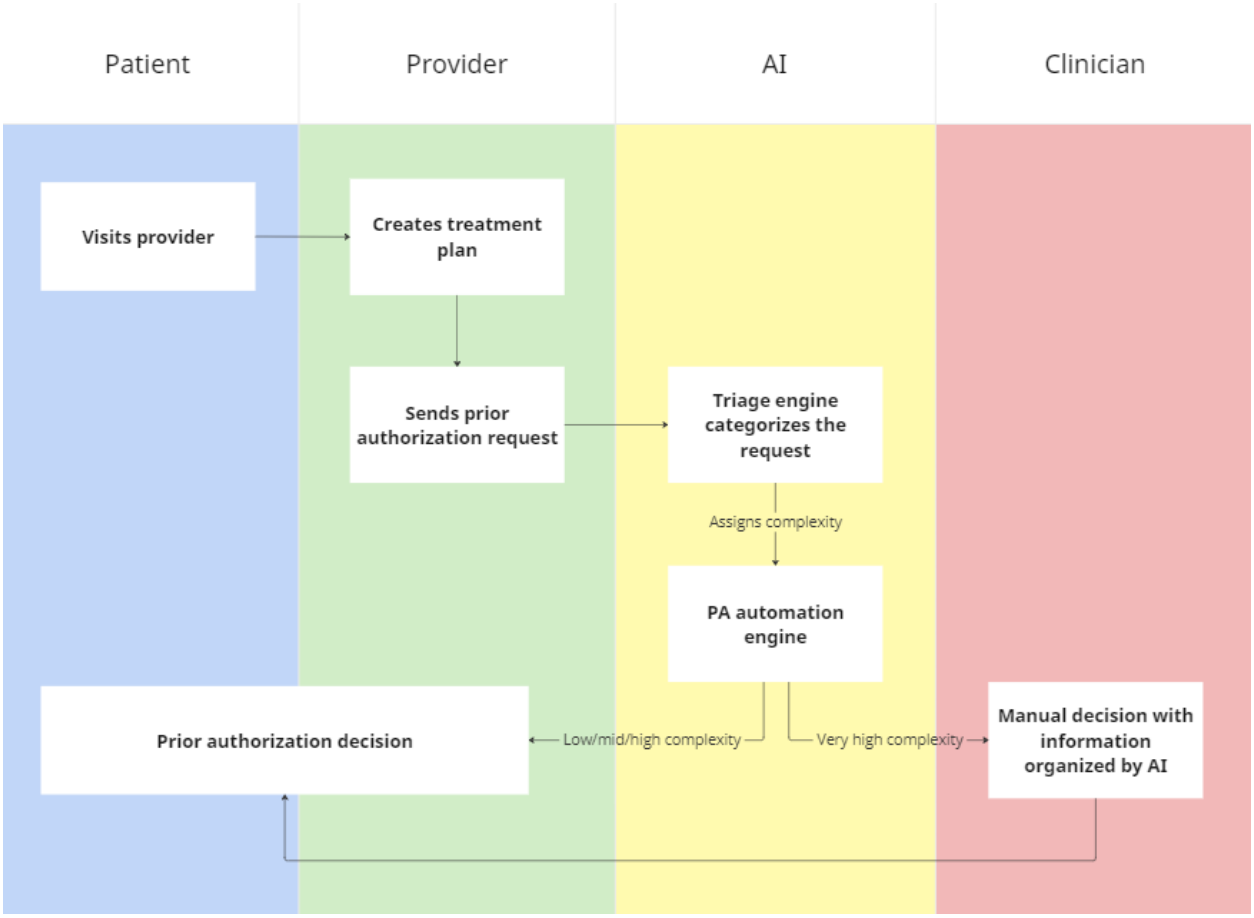
Implementers can add some automation to the existing prior authorization process with no AI assistance for categorization or decision-making. Low complexity authorization may use rule-based decision-making with clinician final approval. Clinicians manually review more difficult authorization cases.

Limitations

The AI process may not balance the interests of patients, providers, and payers. Another AI process for prior authorization seeks to solve this issue. It validates the algorithm from an objective public review and certification against a panel of clinical cases juried by national clinical leaders.

Proprietary implementations are deployed in the marketplace, however prior authentication standards have not been fully tested in large-scale production. The Office of National Coordinator (ONC) has not issued rules for prior authentication use based on Centers for Medicare & Medicaid Services (CMS) rules. The ONC requirement for certified electronic health record technologies (CEHRT) to include electronic prior authorization would be a major market driver for prior authentication deployment and standardization.

Figure 5: Prior Authorization Use Case Swimlane Diagram



Genomics Use Case: Precision Oncology with Genomic Markers

Use Case and Goals

Artificial Intelligence (AI) for precision genomics for cancer treatment planning aims to identify personalized treatments, including clinical trials for cancer patients by integrating their molecular tumor profiles into clinical decision-making. This approach allows selection of an optimal therapy to help maximize patients' survival and quality of life, by delivering the right cancer treatment to the right patient at the right dose and the right time, leading to enhanced treatment efficacy and reduced side effects [143], [161], [162], [163], [164], [165], [166].

AI Algorithm Type

Bayesian/statistical algorithms, Deep learning, natural language processing (NLP). NLP algorithms that match patient molecular and clinical data with available clinical trials based on inclusion and exclusion criteria.

Description

The FDA cleared, tumor profiling test (assay) is a targeted NGS test using DNA isolated from FFPE tumor tissue specimens, and DNA isolated from matched normal blood or saliva specimens, from previously diagnosed cancer patients with solid malignant neoplasms. The test provides information on somatic mutations (point mutations and small insertions and deletions) and microsatellite instability (MSI), tumor mutational burden (TMB) for use by qualified healthcare professionals in accordance with professional guidelines.

The test types can include for example, 1) detection of tumor gene alterations in a broad multi-gene panel that is not conclusive or prescriptive for labeled use of any specific therapeutic product; or 2) a companion diagnostic (CDx) to identify patients who may benefit from treatment with the targeted therapies listed in the Companion Diagnostic Indications table in accordance with the approved therapeutic product labeling.

End Users and Stakeholders

Precision genomics for cancer treatment planning AI tools are used by patients, clinicians (pathologists, oncologists), bioinformaticists, and geneticists. The main stakeholders for these personalized treatment applications include.

- Patients
- Health AI vendor
- NGS sequencing companies
- NGS analysis tool developers
- EHR vendor
- Payers
- Pharmaceutical companies
- Molecular lab testing companies
- Healthcare administrators
- Regulatory bodies (FDA)
- Physicians who treat cancer patients (oncologists, molecular pathologists)
- Groups publishing guidelines, including, Association for Molecular Pathology (AMP), College of American Pathologists (CAP), and National Comprehensive Cancer Network (NCCN).

Model Output / Decisions and Actions Taken

The analysis of sequencing data from the NGS test, from the tumor sample and if applicable the normal sample consists of several steps with different kinds of algorithms. As an example, there are tests that can detect certain mutations in KRAS and NRAS genes to help doctors identify if a person with colorectal cancer may benefit from personalized treatment with approved FDA therapies, ERBITUX (cetuximab) when there is an absence of mutations in codons 12 or 13 of KRAS.

NGS bioinformatics pipeline consists of following steps: 1) Detection and analysis of raw sequencing data resulting in a FASTQ file, 2) Alignment of reads against the reference genome, resulting in variant call files (VCF) for the tumor and normal samples, 3) Somatic mutation analysis is a paired sample variant calling that is performed on tumor samples and their respective matched normal controls to identify single nucleotide variants (SNVs) and indels. These steps involve Bayesian and/or ML algorithmic approaches.
Genomic signature analysis includes for example Microsatellite Instability (MSI) status calling. This test can be based on for example a univariate logistic regression classifier to classify tumors into three categories, and this information can be used to identify treatment.
Matching patients to treatments and trials: Web-based computational platforms can be used by clinicians to automatically match patients' genomic-specific events to approved treatments and clinical trials. NLP-based approaches are being investigated to improve the matching of patients to trials, to derive patient and tumor attributes from EHR and to match the data to therapies and clinical trial eligibility criteria. Annotation that is automated through AI but guided by experts' clinical input.

Interface, Application, & Technological Environment

Several decision support tools integrate the entire or part of the workflow. These applications rely on access to:

- high throughput sequencing machines,
- NGS bioinformatics pipelines hosted on the cloud,
- custom statistical/ML algorithms,
- clinical knowledge bases accessible via web interfaces or databases,
- EHR and
- laboratory information systems for tracking samples.

The application generates a report for the oncologist, who can share it with their patients. This report usually includes a list of variants and recommendations for clinical trials and treatments based on the results. The report content should clearly communicate the association between the genetic variants identified for the patient and the suggested treatment options. The report should be understandable both by the clinicians and the patients. For example, it would be helpful to provide plain language explanation of the concepts to ensure patients can easily understand the recommendations and the basis for that.

Privacy and Security Information

Information about genomics-data ownership and adherence to personal-data-protection legislation need to be properly presented to patients, particularly as patients may receive care from different professionals and institutions along their treatment journey. Additionally, in case of germline testing genomic data can be used to identify relatives which has privacy impact beyond the patient.

Data sources and training

The use case determines the data sources and training methods.

1. NGS pipeline (Variant Calling): The model training processes will use genomic and optionally clinical data.
2. Treatment Recommendations: The model training process will use patient genomic and clinical data.

Additionally, it can include information from clinical knowledge bases, and clinical trials data from https://clinicaltrials.gov/.

Data Flow (Input/Output) / Pre-conditions

The data flow integrates multiple information systems for the precision genomics-enabled cancer treatment planning process.

- A high throughput sequencing machine automatically flows NGS data from a tumor and if applicable a matched normal sample.
- The Bioinformaticist/NGS pipelines should have access to analyze the NGS data, which then flows into a custom software that will annotate the variants.
- Pathologists analyze variants via a software decision support tool and apply filtering and prioritization tools to select a set of actionable variants.
- Next, a software matches the variants with treatment recommendations and generates a report for an oncologist to review the information to make a treatment decision.
- The EHR stores a report for research purposes.

Basic Flow

1. The clinician diagnoses cancer based on a patient's appropriate tests such as imaging and lab tests. The patient provides informed consent for the paired tumor-normal sequence analysis.
2. A sample of the patient's tumor is taken by the provider. The patient provides a blood sample as a source of normal DNA for comprehensive genomic profiling by a diagnostic or in-house lab.
3. Automated protocols extract DNA from tumor and blood samples. Sequence libraries are prepared and captured using hybridization probes based on the NGS target panel.
4. Paired reads are analyzed through a custom bioinformatics pipeline that detects multiple classes of genomic mutations and rearrangements, as discussed above.
5. Genomic signature analysis is performed for example, MSI, TMB. (If applicable)

6. Automated software performs variant/mutation annotation where predicted functional effect and clinical interpretation for each mutation is curated using information from several databases, for example the FDA-recognized Memorial Sloan Kettering's Precision Oncology Knowledge Base (MSK OncoKB, https://www.oncokb.org), and a custom proprietary database,. Results are loaded into a genomic variants database where a medical professional manually reviews for quality and accuracy. Automated software annotates approved results with CDx relevant information merging it with patient demographic information and any additional information provided by the testing lab prior to approval and release by the laboratory director or designee.
7. An AI solution leveraging natural language processing (NLP) techniques takes the individual's molecular profile (tumor genomic profile with annotations; immune profile, such as TMB and MSI status, etc.) and the clinical profile (demographics, clinical presentation, pathological diagnosis, sites of metastatic disease, etc.) and matches it with guideline-recommended therapeutic regimen. This match relies on using an evidence-based annotated database of treatment options. Specific eligibility criteria for these matched therapies can include a certain cancer type, stage setting (metastatic versus non-metastatic) and/or line of therapy, among other criteria.
8. This report is reviewed and signed off by the laboratory director or designee for accuracy.
9. The oncologist reviews the report and discusses it with the patient to select a personalized treatment, such as a clinical trial and/or an approved or guideline-recommended therapeutic regimen intervention(s), in the context of supporting evidence, potential benefits and possible risks.
10. Optionally, genomic alterations are reported in the EMR, transmitted to an institutional database that facilitates automated clinical trial matching and automatically uploaded to a portal (e.g. cBioPortal) for data mining and interpretation.

Alternative Flow

In the absence of an AI solution to match a patient's profile to available therapies and trials, the clinician can manually search for the therapies and clinical trials based on the patient's profile. For example, the clinician can look for trials that are accepting enrollment on clinicaltrials.org website (or an institution-provided clinical trials database) to identify trials that match the patient's profile.

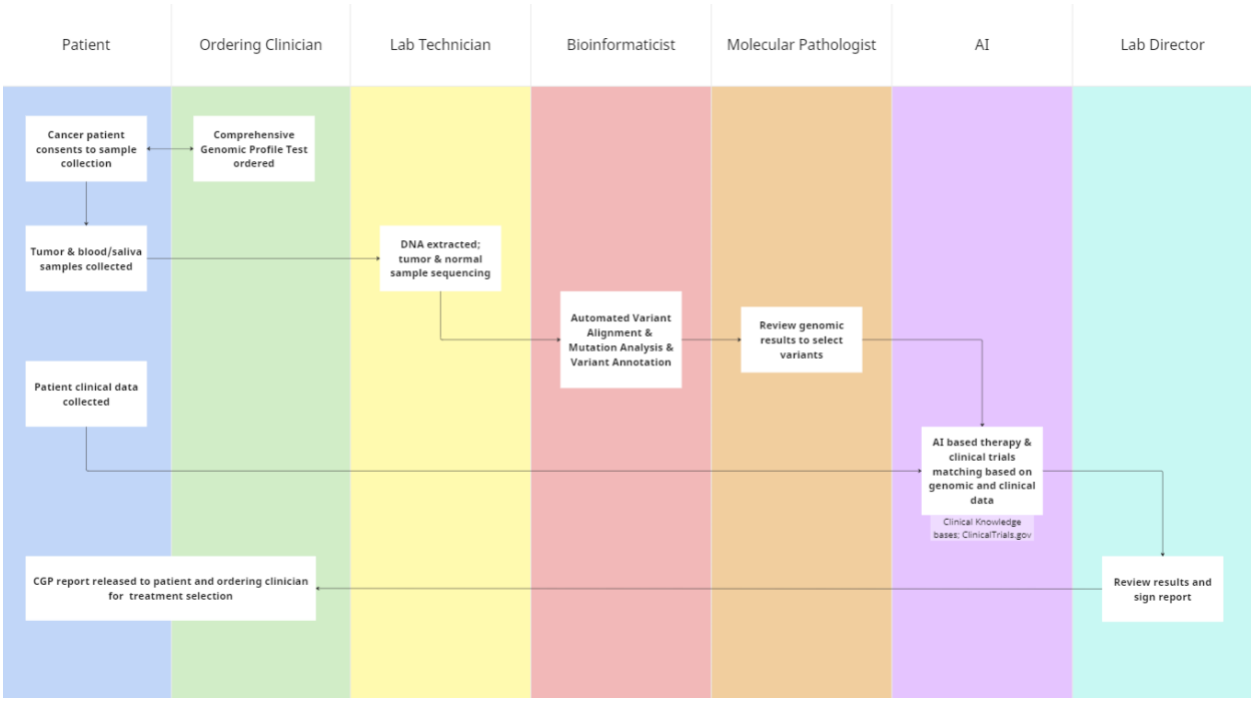
Limitations

AI cancer models of today have a strong emphasis on image and -omics data, however one of the richest data sources is the EHR which remains hugely underutilized. Reasons for this include records being unstructured with high levels of noise, sparseness, and inconsistencies, requiring dedicated curation and data cleaning. These challenges are being actively addressed by standards such as the Observational Medical Outcomes Partnership Common Data Model, which is focused on restructuring patient data into easy-to-use databases with standardized disease codes and harmonized vocabulary.

Clinical trial enrollment has a number of barriers to trial enrolment, for example, clinical trials are performed only at specific locations, which limits their access to some patients and populations. AI and digital health solutions can have an impact on the barriers related to clinical trials.

The safety risk of a false negative would be much higher as it would mean missing a cancer diagnosis. The risk of a false positive is anxiety, costs associated with additional testing and in some cases could result in unnecessary procedures for a patient. To mitigate this risk, these test results can be combined with imaging/ lab tests as applicable for the specific cancer type.

Figure 6: Genomics Use Case Swimlane Diagram



Appendix 2: Expanded AI Lifecycle Framework

This AI Lifecycle Framework is intended to guide a diversity of responsible parties and stakeholders involved in bringing AI solutions into routine use in health systems: AI developers, product and design teams, data scientists, engineers, researchers, health system leaders, clinicians and other healthcare providers, payers, patients, and other end users. Each stage of the framework is grounded in the needs and considerations of health systems, which represent the primary customer and deployment setting of health AI solutions. Still, the questions and recommendations accompanying each stage are intended to inform best practices and actions for all responsible parties.

In many cases, responsible parties span organizations and roles. We will use the phrase *developer team* to refer to the responsible parties primarily involved in the AI solution development process; they may consist of data scientists, software engineers, healthcare providers, or a subset of those. Similarly, the *implementer team* comprises responsible parties involved in implementing and integrating an AI solution in health system workflows, including data scientists, data engineers, human factors and behavioral science professionals, user experience and user interface designers, health system leadership, etc. The composition of each team may vary depending on the use case and setting. For example, the developer team may be employed by a software company, while the implementer team may work for a health system. In other cases, the developer and implementer teams may all be health system employees. When the developer team is an employee of a software company making a new AI solution, they often are involved in the early stages of problem identification, planning, and design, working alongside the implementer team. They may bring insights from those processes into solution development and commercialization.

The 6 Stages of the AI Lifecycle described here are as follows:

- Stage I: Define Problem & Plan
- Stage II: Design the AI System
- Stage III: Engineer the AI Solution
- Stage IV: Assess
- Stage V: Pilot
- Stage VI: Deploy & Monitor

The Lifecycle borrows aspects from agile software development methodology, and as such is intended to support flexibility when moving between Stages, especially Stages II-IV. For example, suppose aspects of AI solution design require refinement based on information gained during the assessment stage. That process should be undertaken, with potential follow-on refinements in later Stages as needed.

The remainder of the document describes each *Stage* as a series of *Steps*. Responsible AI Checkpoints can be added to different steps of the lifecycle but generally would be before the AI system is piloted in a real world setting (before Stage V) and before AI is broadly used on a

general population (before Stage VI). Responsible AI checkpoints should be added at frequent intervals in the Deploy & Monitor Stage (Stage IV) to assess drift, safety risks, impact, and adoption.

Stage I: Define the Problem and Plan

Summary

Healthcare is riddled with "solutions" in search of problems. Responsible innovation requires a clear understanding of the specific issue AI is intended to solve, which will drive the intended purpose of the tool. The intended purpose a) defines the scope of verification & validation activities and b) allows reasonable delineation between user responsibility and vendor responsibility. First, however, an upfront investment of time and effort is needed to map root causes and understand the specific needs of those experiencing the problem(s) via primary research. Only *after* a problem is clearly defined can health systems and developers begin brainstorming potential solutions, and whether AI is an appropriate tool. The problem, its setting, and the personnel involved comprise a *use case* where AI may be a solution. Selecting a solution from a potential set depends on the potential for a given solution to positively impact relevant patient and clinician outcomes, stakeholder engagement, legal implications, estimated return on investment, and health system resource allocation and prioritization processes.

For developer teams, Stage I must involve market research to identify the problems and challenges that potential customers grapple with across healthcare organizations. For implementer teams, Stage I focuses on understanding a specific problem and potential return on investment (ROI) and return on health (ROH), given options for the solution and the intended future state and associated business requirements. Developer teams may participate actively in this stage by sharing information around feasibility and by gathering information that may inform the design of the AI solution (Stage II).

Stage I Steps

1. Engage stakeholders to define the problem and perform root-cause analysis
    - What is current state?
    - What is the problem to solve? Does it necessarily require an AI solution?
    - What is the intended use of the proposed AI solution?
    - What are the ethical considerations relevant to the proposed AI solution?
        - Differences in expectation or perceived impact across stakeholder groups
        - Is there potential for patient subgroups to be differentially affected?
        - What are the potential unintended consequences of using the proposed AI solution?

2. Identify solution and plan future state
- What is the downstream impact of the proposed AI solution?
- How will the clinical and/or business workflow change by introducing an AI solution?
    - What is the existing workflow? What is the capacity of the team executing the workflow?
    - Where will the AI solution be inserted into the existing workflow (or new workflow, if one does not exist)?
    - How might the AI system introduce risks and hinder current processes, i.e., patient-clinician interaction? How will those risks be mitigated? (early draft risk management plan)
- How will we know we have solved the problem?
    - What are the metrics and key performance indicators we should use to measure the impact of the AI solution?
    - If the AI tool is intended for use in clinical decision-making, will it lead to better outcomes than the current standard of care?
    - What is the timeline in which we expect to see the desired outcomes?

3. Gather business requirements
- Stakeholder Research
    - What are all the types of stakeholders who may be impacted by the proposed AI solution? Are their needs and viewpoints incorporated into the problem and solution identification steps?
    - How will stakeholders be engaged in using the proposed AI solution?
    - What will end users need to put trust in the AI solution and its output?
- Include Legal & Policy Considerations
    - What are the liability risks for clinicians and health systems using the proposed AI system?
    - What health system policies apply to the proposed AI solution regarding approval for use, monitoring, retiring etc.?
    -

4. Assess feasibility, potential for impact, and prioritization
- What are the estimated ROI and ROH if the proposed AI solution is implemented? This may refer to financial costs/profit, patient outcomes, and health system personnel efficiency gains or losses, risks, and satisfaction.
- Does the proposed AI solution align with the strategic initiatives of the health system?
- What resources are available for AI system deployment, and what processes are used to prioritize resource allocation amongst many possible AI systems?

5. Make procure/build/partner decision

Based on the answers to the questions and results of analyses in steps 2-4, for each potential AI solution proposed for the specific use case, the implementer team will decide whether to *procure* an existing AI solution, *build* an AI solution "in house" or *partner* with a third party (e.g. a software company) to develop a bespoke AI solution specific to the health system where it will be deployed. Procurement includes buying commercially available solutions from third parties or adopting solutions made publicly available via research literature and/or code repositories. Agreements are put in place around the responsibilities of different parties in the subsequent stages of the lifecycle.

Stage I: Decision Point

This stage culminates in a decision by the implementer team to build, procure, or partner to apply an AI solution to the identified problem, or to use an externally validated, open source AI model. Alternatively, they may determine that an AI tool is not required to solve the identified problem.

In the case that an AI solution is deemed necessary, organizational maturity, resource availability, and funding are primary considerations. Specifically, the health system where the AI solution will be implemented requires personnel, financial, and material resources to train users, evaluate the AI system, develop workflows (where applicable), monitor, and prospectively evaluate impact. If a developer team participates in this stage by engaging with various stakeholders to gather business requirements (potentially across multiple implementer organizations), they will determine the go-to-market strategy for their product and whether they will partner with an implementer team to develop an AI solution.

Stage II: Design the AI System

Summary

After defining the problem and the proposed solution, and after the implementer team has decided whether to procure, build, or partner in developing the solution, Stage II begins the solution design process. This involves capturing the solution's technical requirements, the intended scope of the solution, the proposed system workflow, and deployment strategy. The design of the solution is informed by the business requirements of the health system(s) and the needs of end users where the AI solution will be implemented.

Developer teams working at AI software development companies may participate in the design stage with many implementing organizations (their customers) in a pre-market phase. This has the potential to result in a more robust product suitable for multiple future customers. The implementer team is primarily responsible for capturing details to design the system workflow, organizing requirements for monitoring and reporting, and designing a deployment strategy.

Stage II Steps

1. Select/understand model task and architecture
    - The decision to procure, build, or partner informs the need to either understand the AI solution's model task and architecture, verifying that it is appropriate for the use case or to select the appropriate task and architecture informed by the use case and business requirements. Model task refers to the type of prediction or classification made by the AI solution, e.g., is it a risk prediction task where the output is a probability of a specific future event occurring for a patient? Is it a classification task where the model will output a binary "yes" or "no" flag about patient state at the time of inference? Model architecture refers to the type of model learning algorithm and structure of the resulting model that will be used. Examples include decision trees and deep neural nets.

2. Capture design, data, and technical requirements or determine the best solution to meet business requirements
    - If procuring an AI solution, this step involves the implementer team selecting among the available options to identify the best solution that satisfies business requirements. Business requirements include the size of the institution where the solution will be deployed, the budget available for purchase and implementation, and stakeholder needs as surfaced during Stage I.
    - If building or partnering with a third party to build an AI solution, this step involves further stakeholder interviews by the developer team to capture design and technical requirements that the solution must meet.

3. Design solution application and system workflow
    - If procuring an AI solution, review the AI solution user interface design to verify its appropriateness. If building or partnering, design the user interface and actions for AI solution.
    - Design the workflow, including the intervention that will be taken based on AI solution output
    - Plan alternative workflow(s) for downtime and potential sunsetting
    - Use human-centered design principles and processes
        - Understand the environment of use by observing end users
        - Identify intended users and user needs by interviewing end users
        - Iterative, participatory design with users, identifying design requirements and anticipated risks

4. Design deployment strategy with end users
    - The implementer team will design how the AI solution will be deployed (e.g. using on-premise or cloud compute; managed by what team etc.), including the infrastructure and resources that will be needed.

    - Workflow Integration – Process mapping (Swimlane Workflow Diagram vs. Current State Workflow Diagram)
    - Access to Transparency Information
- Plan how the AI solution deployment will scale, including the resources that will be needed, considering and documenting in a risk management plan, issues and risks that may be encountered in real-world health system settings.

5. Design risk management, monitoring and reporting plan
- Design an initial version of a risk mitigation plan for potential risks and challenges associated with the deployment of the AI solution, such as bias, fairness, safety, and security risks, to be revised and implemented during Assess, Pilot and Deployment stages.
- Based on AI solution application and system workflow design, identify AI solution outputs and corresponding health system outcomes that should be monitored after deployment.
    - Data Integrity: how will accuracy, completeness, and quality of data be assessed over time?
    - Impact/Outcome Measurement Plan
        - Localization: How will AI solution outputs and impact be evaluated and tuned to "localized" data upon which the solution will be applied?
- Feedback Framework: how will stakeholders, particularly end-users, communicate questions or concerns to trigger possible re-evaluation?
- Design a report that will summarize monitoring results for AI solution outputs and health system outcomes at appropriate timepoints e.g. weekly, monthly, quarterly (depending on use case).

Stage II Decision Point

This stage culminates in a design for the system application, corresponding workflow, and deployment strategy. These designs will be used as the basis for engineering the AI solution (when applicable) or to determine with the developer team how a current commercially available solution should be adapted. The implementer team will determine the strategy for how the AI solution should be deployed.

Stage III: Engineer the AI Solution

Summary

The engineering stage aims to create an AI solution that can accurately predict or classify data and develop the interface for the model, as defined during the Design stage. This stage also ensures that AI solution deployment can be operationalized and that adequate planning is

completed prior to deployment. In cases of externally developed AI solutions, the developer should provide expertise in collaboration with the implementer, who ensures that the AI solution meets its intended purpose via risk-benefit analysis before and after deployment.

Data access, preparation, and management are the processes of obtaining, cleaning, and organizing data to be operationalized and used for AI solution development and analysis. It is a critical part of the AI lifecycle, ensuring that data are accurate, reliable, and accessible. Data engineers use a variety of tools and techniques to transform raw data into useful information that supports downstream use cases. They may use programming languages to extract data from databases or data visualization tools to create charts and graphs.

Model training and tuning are critical parts of building an AI solution and are an iterative process. The quality of the model will depend not only on the quality of the data but also on the selection of the algorithm and the training process. Through an agile and iterative development process, data scientists use statistics, machine learning, deep learning, natural language processing, computer vision, forecasting, optimization, and other techniques to understand the data, select an approach, train the model, and evaluate its performance.

Much of the engineering process is led by the developer team. The implementer team is responsible for validating the appropriateness and feasibility of the processes for data access, preparation, and management, as well as model training and tuning. This involves close collaboration with the developer team.

Stage III Steps

1. Access data
    - Identify data sources, extract data from those sources, and load data into a data warehouse or data lake. Types of data may include synthetic data, an extract/copy of real data, or live-streaming data.
    - Accessing high quality, relevant data is often a significant challenge. Data may be fragmented across sources, inconsistently captured, or of low quality, which may affect AI solution training, performance, and reliability.

2. Prepare data
    - Clean, transform, and organize data to serve the highest quality data. Data preparation accounts for nearly 80% of AI development efforts and should follow well-accepted interoperable standards for data transfer (such as FHIR) and storage (common data models such as OMOP, PCORnet, and i2b2 are among many options), when applicable.

3. Develop data management plan

- Catalog data assets and classify data lineage to ensure transparency and trust in data sources used for model training and evaluation.
- Implement a system to log data access and updates.

4. Train and tune model

- When built in-house or through collaboration with an industry partner, a model is trained on a set of data. In the case of supervised machine learning, the data set is labeled with the desired output, and the model then learns to associate the input data with the output data. Once the model is trained, it can be used to make predictions on new data (prospective data). In the case of unsupervised machine learning, or when engineering generative AI solutions, other training techniques may be used.
- Multiple model learning algorithms may be used and compared as appropriate. In most cases, the model is developed on retrospective data.
- Tuning a model may be appropriate for AI solutions that are built in house or in partnership with a vendor. Tuning involves modifying the values of model hyperparameters to maximize model performance.

Stage III Decision Point

This stage culminates in a quality-assured dataset with documentation supporting lineage, and a fully-developed model with validated outputs and, where possible, impact. (In certain instances, the impact of the AI solution can only be assessed in a real-world setting.) With the model in hand, the team may advance to the next stage for a business decision of whether to deploy the AI solution into the health system, or, when applicable, return to the Stage II to refine the design of the AI solution or corresponding workflow. When partnering with a third party, or procuring a solution, this may affect the degree of solution customization that is possible, and may also require planning around intellectual property rights.

Stage IV: Assess

Summary

This stage involves a series of assessments to determine whether to proceed with a pilot of the AI system. When AI-enabled technologies are acquired from a third party, local validation and installation qualification need to be conducted first, prior to the assessment of the AI system. A change management plan should be in place to delineate who, between the developer and implementer, is responsible for performing these duties. This is followed by a prospective, silent evaluation and the establishment of a risk management plan. Such a plan ought to manage contingencies related to poor performance of the AI solution (e.g., biased outputs), changes in the deployment environment (e.g., changes in outcome prevalence and data drift), and unanticipated misuse of the AI solution. These steps are followed by end user training and usefulness testing, along with a review to ensure compliance with applicable healthcare

standards and regulations prior to piloting and deployment. A business/clinical owner should be defined, who will be accountable for ensuring that the AI solution is tested and that personnel are trained, eliciting their feedback.

These processes are primarily the purview of the implementer team. The developer team may also use the results of the Assess stage to quantify their product's functionality and efficacy, which may inform iterative product development.

Stage IV Steps

1. Conduct installation qualification (when applicable)
    - This step is applicable to AI solutions procured from third parties to verify they are correctly installed. This can be performed by the implementer team, or the health system can rely on the third party the solution was purchased from to perform this step.
    - Assess the technical correctness of the AI model's installation
    - Ensure the correct installation of software and hardware
    - Document the installation process for regulatory and accountability purposes

2. Validate local system performance (when applicable)
    - This step is applicable to existing AI solutions procured from a third party (e.g. commercially available solutions, pre-trained AI models in public repositories etc.).
    - Develop a representative test dataset from the specific deployment service area
    - Verify the integrity of data inputs and outputs from the AI solution using the test dataset
    - Confirm compliance with operational specifications and requirements
    - Deploy the AI system in a local, simulated environment
    - Assess system performance, potential impact and generalizability
    - This validation should be performed by the developer and implementer teams; documentation that the AI solution performs as expected should be collected.

3. Execute prospective, silent evaluation
    - A silent evaluation involves generating AI system output using production data, but not displaying that output to personnel who would take action as part of the system workflow.
    - Prospective evaluation against the live data source used in production, monitoring data quality and AI system behavior.
    - Operationalize & optimize: Embed AI solution into an operational system to evaluate output and performance
        - Generate evidence of the AI solution's potential effectiveness, safety and fairness via performance testing and unit testing.
            - Independent evaluation or external validation is often performed as a common practice to test the robustness of technology
    - Pre-pilot planning
        - Define pilot scope
        - Classify intended impact (noting whether the output will "touch" patients)

- Establish a change management plan among developer and implementer teams, e.g. determine which party will be responsible for performance monitoring, updates of model and interface, etc. The implementer team will be responsible for assessing whether the solution meets impact, safety and other business related success criteria
- Refine success criteria from Stage I as needed (include end-user acceptance criteria)

4. Establish risk management plan
- Risk mitigation planning: This involves planning for potential risks and challenges associated with the deployment of the AI solution, such as bias, fairness, safety, and security risks, based on stakeholder input on workflow design and testing (see step 6, below).
    - Create a deployment bias evaluation and management plan
    - Create an incidence change detection and response plan
        - Establish thresholds for data/feature/concept drift, bias in subpopulations, performance drop, outages, bugs in user-facing code, etc. that would result in decommissioning AI solution or trigger further investigation.
        - Corrective and preventive action (CAPA) plan
            - Capture safety, bias, usability and other issues and define strategies for prompt mitigation
            - Make Contingency management, decommissioning, and rollback/backup plans
    - Create a misuse mitigation plan

5. Train end users
- Train a sample of end users on how to use the AI solution in their work
- Prepare training materials and documentation of AI solution design, safety, risks, intended purpose, etc.
- Prepare training materials on how to identify and report issues with the AI solution.

6. Test usefulness
- Evaluate
    - Survey and/or interview a sample of end users to collect feedback on system safety, efficiency and effectiveness for its intended use
- Usability testing (formative and summative) with end users
    - Evaluate whether intended users can complete realistic tasks efficiently and effectively; identify areas of confusion when users interact with the system
    - Evaluate user satisfaction with the system with interviews and/or standardized surveys

- Refine the system design (e.g., interface, workflows) based on task successes/failures and user feedback
- Documentation of meeting requirements and success criteria
- Refinement of training material, transparency material, CAPA and risk management plan as needed, based on user feedback.

7. Ensure compliance with applicable healthcare regulations and standards
- Investigate compliance of AI solution with applicable healthcare regulations and standards (government regulations, HIPAA compliance, etc.)

Stage IV Decision Point

This stage culminates in a business decision to deploy the AI application (or not) as a pilot. The decision to pilot is accompanied by approved implementation, measurement, and mitigation plans, as well as pilot user training, *prior to* deployment.

Stage V: Pilot

Summary

The pilot is the first real-world use of the AI solution that informs large-scale deployment plans. Prior to the general deployment of an AI system, careful review and consideration must be made by the health system to decide whether or not to deploy an AI model into production. For a Go/No-Go Decision to be made, success criteria are reviewed to inform the decision on whether to deploy the AI system, based on the results of a pilot. Some common criteria include the AI solution's accuracy, reliability, interpretability, feasibility, user acceptance, cost, and alignment with the organization's values and goals. This process is primarily undertaken by the implementer team. For the developer team, this stage can identify settings where their product has the desired impact and is useful.

Stage V Steps

1. Assess Real-World Impact
    - Small Scale Safety & Utility – In a real-world setting, evaluate the effectiveness of AI solution-guided decisions and impact (DECIDE-AI) while monitoring and reporting according to Monitoring and Reporting Plan (Stage II), using statistical methods
        - Examples of study design approaches: stepped wedge design, A/B testing via randomization
    - Usability and impact on workflow - evaluate how the AI integrates in clinical and/or operational workflows and any unanticipated issues from the implementation of the AI solution. Gather end user feedback on the tool and any challenges or barriers to use.

Evaluate end user trust in the system. Consider if any design changes are needed to optimize use of the tool by end users.

2. Execute and update risk management plan
   - Use the risk management plan created in Stage IV, updating as needed to support new scenarios observed during pilot.

3. Educate and train users on AI application and reporting
   - Training material and documentation: Information on the model's intended use, architecture, training methodology, performance, and known limitations and safety risks is important to formalize to ensure understanding, correct use, reproducibility, and troubleshooting. Establish readiness for audit by compiling all AI system documentation, including the monitoring and reporting plan defined in Stage II, as well as training materials.
   - Share training material and documentation with end users, and offer training. Conduct stakeholder interviews and feedback sessions around usefulness and adoption of the AI solution, to understand successes and limitations.
   - Put Infrastructure in place to support monitoring and reporting plan from Stage II. This is typically managed by the implementing team.

4. Assess Usefulness and Adoption
   - Evaluate workflow integration, end user acceptance, and potential downstream impacts of the AI solution.

Stage V Decision Point

At the conclusion of Stage V, readiness for larger-scale deployment and monitoring in Stage VI will be established (or not). After the pilot stage, integration testing and optimization is performed before larger-scale deployment (Stage VI).

Stage VI: Deploy & Monitor

Summary

Deployment is the process of making the AI solution and system broadly available to the health system or relevant specialty. Once deployed by the implementer team, the AI solution is often handed over to a model operations team (when available) to provide ongoing monitoring, retraining, and governance of models to ensure peak performance and that decisions are transparent. The developer team may also benefit from regular reporting out of deployment

results to inform their product strategy, and to identify monitoring and reporting requirements that their product needs to satisfy.

Stage VI Steps

1. Deploy at a larger scale on a general population
    - Following the deployment plan developed in Stage II (and then stress-tested on a small scale during Stage V), deploy the AI system on the general eligible health system population.

2. Audit AI system to inform whether to maintain, refine or sunset
    - Review AI system performance and impact using the monitoring and reporting plan developed in Stage II and adapted in Stage V on a consistent cadence (e.g. quarterly, yearly) to decide whether to maintain the system as deployed, refine aspects of the deployed system, or sunset ("turn off") the system.
    - Review end users feedback of the tool periodically and monitor for issues that emerge from use of the AI over time. Evaluate if design changes are needed as the system and workflows evolve over time.
    - If refinement or sunsetting are recommended, ModelOps to collaborate with Business Operations, Data Engineering, and Data Science teams to develop and execute refinement or sunsetting procedures.

3. Conduct ongoing risk management
    - Use risk management plan created in Stage IV, updating as needed to support new scenarios observed during deployment on the larger eligible population.

Stage VI Decision Point

This stage culminates in a successfully deployed AI system with ongoing monitoring. If and when AI solution performance drifts or deviates, the AI solution may be revised, possibly returning to Stage II or Stage III, or the AI system may be decommissioned entirely.

Governance

Governance by implementer and developer teams is important for patient safety and ensuring the trustworthiness of AI solutions. It is helpful to assign roles and responsibilities for the initiative early in the planning process, using the RACI Matrix:

- Responsible: The person or team who is accountable for completing the task.
- Accountable: The person or team who has the final say on the task.
- Consulted: The person or team who is asked for input on the task.
- Informed: The person or team who is kept updated on the progress of the task.

Stage I concludes with a decision to procure, build, or partner with a third party to develop the AI Solution. In all three scenarios, following robust governance principles during the design phase is key: safeguarding autonomy, promoting human well-being and safety, fostering transparency, ensuring accountability, encouraging inclusiveness and fairness, and sustaining a responsive AI ecosystem (WHO, 2021).

If the decision is made to procure or partner with a third party, additional governance considerations are important:

- Third party transparency when purchasing a commercially available AI solution or partnering to develop an in house solution: Building trust between the implementer team and developer team is crucial. This can be achieved by defining clear procurement requirements, evaluation criteria, and contractual terms.
- Explainability: A key aspect of AI transparency, should be an integral part of the entire machine learning (ML) workflow – from data collection and processing to model training, evaluation, and deployment (Lakshmanan, 2021). In essence, explainable AI provides understandable reasons for model decisions, enhancing trust in AI tools.
- Localization: Ensure that the procured AI model is effective, reliable, and fair when used in its specific intended environment.
- It is also important to ensure that guardrails are well defined and documented for who is approved to access a given dataset or data type, and for what purposes. Good data governance will also foster data democratization. In terms of accountability, establishing standard operating procedures and specifying roles and responsibilities are key, as are training and qualifying those involved with development, deployment and use. It is vital to have qualified health care personnel in the loop, and patient/caregiver representatives when they may be end users. Lastly,accountability of local governance of implementer and developer organizations, to provide oversight in governing and assuring health AI solutions for their trustworthy use (risk management, monitoring performance and outcomes, etc.) is crucial.

Appendix 3: Privacy and Cybersecurity Profile

Introduction

Artificial Intelligence (AI) technology has tremendous potential to transform healthcare by improving clinical decision-making, personalizing patient care, and optimizing operational workflows. However, it also raises significant legal and ethical risks around data privacy and cybersecurity, algorithmic bias, safety, transparency, and the doctor-patient relationship. As AI becomes more widely adopted in healthcare, we must ensure it aligns with human values and enhances, rather than replaces, human skills and judgment. This will require building thoughtful governance frameworks and multidisciplinary collaboration between technologists, healthcare professionals, ethicists, and patients. The goal should be leveraging the respective strengths of both humans and machines to provide the best possible care and outcomes for all.

The Coalition for Health AI (CHAI) is a community of academic health systems, organizations, and expert practitioners of AI and data science. Recognizing the potential of AI technology to transform healthcare, as well as the risks inherent in using and developing these technologies, these members came together to provide guidelines regarding an ever-evolving landscape of health AI tools to enable the fair, transparent, and safe implementation of AI in healthcare. In 2022, CHAI published a "Blueprint for Trustworthy AI in Healthcare," a first step towards developing implementation guidance on trustworthy AI in healthcare. In 2023, CHAI engaged a wide range of stakeholders from government, academia, and industry to take the next step in operationalizing the principles set forth in the Blueprint, including a Responsible AI Guide and Checklist that set forth key considerations and evaluation criteria throughout the AI lifecycle that developers and deployers of AI in healthcare should incorporate into workflows. One of the CHAI work groups initiated as part of this work was the Privacy and Cybersecurity Work Group. This Healthcare AI Privacy and Cybersecurity Framework Profile ("CHAI Profile") is an output from this Work Group. The CHAI Profile, created through collaboration with a diverse range of stakeholders, uses the NIST Privacy v1.0 and Cybersecurity v1.1 Frameworks to provide a prioritized, risk-based approach to address privacy and cybersecurity issues that are unique to the use of AI in healthcare. The Frameworks present a variety of risk management outcomes organizations may wish to achieve, and the CHAI Profile tailors and prioritizes those outcomes for healthcare AI objectives. Profile tailors and prioritizes those outcomes for healthcare AI objectives.

Purpose

The CHAI Profile provides voluntary guidance to help organizations manage privacy and cybersecurity risks for organizations that use AI capabilities to support healthcare research and operations. The CHAI Profile helps organizations prioritize privacy and cybersecurity capabilities based on their Healthcare AI Priorities, which can inform decision making. The CHAI Profile is intended to aid organizations with organizing and communicating their existing and future privacy and cybersecurity activities, practices, policies, and guidance. Organizations should consider their own obligations, operating environment, and Healthcare AI Priorities when prioritizing and implementing privacy and cybersecurity capabilities and controls.

Healthcare organizations using AI use the CHAI Profile to:

- Understand privacy and cybersecurity considerations that are relevant to the use of AI in healthcare

- Assess current organizational privacy and cybersecurity practices to identify gaps and areas of improvement for existing practices or infrastructure
- Develop individualized organizational Current (As-Is) and Target (To-Be) Profiles
- Prioritize investments in privacy and cybersecurity capabilities aligned to the PF and CSF Subcategories identified as most important to support organizational Healthcare AI Priorities
- Understand the relationship between privacy and cybersecurity risk management

Scope

The CHAI Profile focuses on privacy and cybersecurity risks that are unique to AI tools and applications throughout the healthcare delivery system. As noted in the NIST AI Risk Management Framework: "Like safety and security, specific technical features of an AI system may promote or reduce privacy. AI systems can also present new risks to privacy by allowing inference to identify individuals or previously private information about individuals." It is these types of risks that we will focus on in the CHAI Profile. Other types of risks with AI are addressed in other CHAI resources.

Audience

The intended audience for the CHAI Profile includes organizations across public and private sectors who use AI capabilities in healthcare. The CHAI Profile can be used by organizations to identify and communicate privacy and cybersecurity expectations with internal and external parties. The CHAI Profile can also be used by organizational leadership to generate priorities tailored to the operational aspects of the organization.

Document Structure

The remainder of the CHAI Profile contains the following content:

- The Promise of AI in Healthcare: Discusses examples of applications of AI in healthcare activities
- Privacy and Cybersecurity Risk Management: Discusses the relationship between privacy and cybersecurity
- Overview of Privacy and Cybersecurity Risk in Healthcare AI: Provides an overview of the privacy and cybersecurity considerations that arise when using AI in healthcare activities
- Profile Development Approach: Describes how the CHAI Profile was developed
- Overview of the NIST Frameworks: Introduces the NIST Privacy and Cybersecurity Frameworks and the elements used to create the CHAI Profile.
- Summary of Healthcare AI Priorities: Describes the healthcare AI priorities around which the CHAI Profile is oriented
- Contents and Use of the CHAI Profile: Explains the type of information provided in the CHAI Profile, helps practitioners understand how to adapt and apply the CHAI Profile in their organization, and provides a table to show the alignment of Healthcare AI Priorities with prioritized Subcategories from the NIST Privacy and Cybersecurity Frameworks

The Promise of AI in Healthcare

AI is already making significant strides in healthcare. Some of its current and potential applications include:

- Disease Identification and Diagnosis: AI algorithms analyze medical images to detect diseases such as cancer at early stages.
- Treatment Personalization: AI can analyze genetic information to recommend personalized treatment plans.
- Drug Discovery and Development: AI aids in predicting how different drugs can treat diseases, speeding up the drug development process.
- Operational Automation: AI streamlines hospital operations from appointment scheduling to patient flow optimization.

- Expand Remote Patient Monitoring: Using wearables and other devices to monitor patients in real-time, predicting potential health issues before they become severe.
- Enhance Virtual Health Assistants: Improving patient engagement and adherence to treatment through AI-powered virtual assistants.

While AI holds immense promise in healthcare, its successful integration must address privacy, security, and ethical concerns to ensure that patient data is protected and used responsibly. There are other risks with AI, but the CHAI Profile focuses on security and privacy risks that are unique to AI tools and applications in healthcare. As noted in the NIST AI Risk Management Framework relating to privacy: "Like safety and security, specific technical features of an AI system may promote or reduce privacy. AI systems can also present new risks to privacy by allowing inference to identify individuals or previously private information about individuals." The CHAI Profile focuses on these risks.

Understanding the unique risks that AI poses to the privacy and security of health data begins with understanding key terms and how AI may be "unlocked" within the healthcare industry. It is incumbent upon healthcare professionals to be aware of the potential AI brings to improving patient care and healthcare operations and assist in the development and deployment of these applications in an efficient, safe, and effective manner.

AI is a broad concept that refers, generally, to the use of machines, software, or systems that can mimic or simulate human intelligence and cognitive functions. It encompasses a wide range of techniques, including natural language processing, machine learning, computer vision, robotics, and more. AI systems can be designed to perform tasks such as problem-solving, reasoning, planning, learning, perception, and language understanding. When we say "AI" in the CHAI Profile, we are talking broadly about any AI methods or tools that may be involved in a healthcare application.

Privacy and Cybersecurity Risk Management

Relationship Between Privacy and Cybersecurity

Cybersecurity and privacy are independent and separate disciplines. However, as shown by the Venn diagram in Fig. 1, some of their objectives do overlap and are complementary. Cybersecurity programs are responsible for protecting information and systems from unauthorized access, use, disclosure, disruption, modification, or destruction (i.e., unauthorized system activity or behavior) to provide confidentiality, integrity, and availability as well as ensuring organizations comply with applicable cybersecurity requirements. Privacy programs are responsible for managing the risks to individuals associated with data processing throughout the information lifecycle[1] to provide predictability, manageability, and disassociability[2] as well as ensuring organizations comply with applicable privacy requirements. Fig. 1 illustrates this relationship between cybersecurity and privacy risks, showing both where they overlap and where they are distinct.

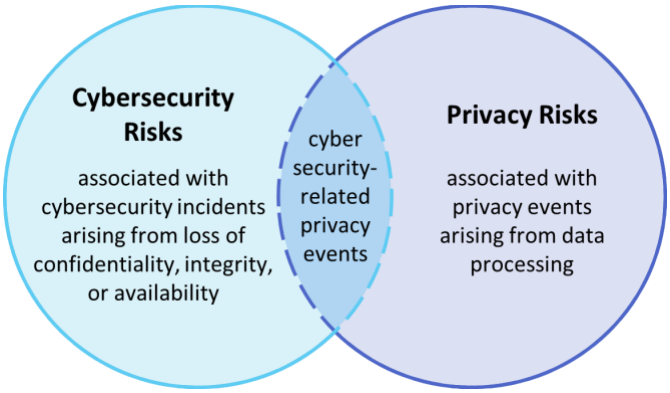


Fig. 1. Cybersecurity and Privacy Risk Relationship (from the NIST Privacy Framework)

While the overlap between cybersecurity and privacy risk management is important, the distinction between the two is also critical to understand. Managing cybersecurity risk contributes to managing privacy risk (e.g., controlling access to data protects against privacy breaches by limiting who can access data and the actions they can perform), but managing cybersecurity risk alone is not sufficient for managing privacy risk, as permitted data processing activities can introduce privacy risks that are unrelated to cybersecurity incidents. Some data processing activities and technologies inherently introduce privacy risk but may be necessary for valid business purposes. These privacy risks must be managed when they arise.

Overview of Privacy and Cybersecurity Risk in Healthcare AI

In the healthcare sector, especially with the advent and rapid integration of AI, these intersections between cybersecurity and privacy have become even more pronounced. AI-driven healthcare solutions often require processing substantial amounts of sensitive patient data — ranging from medical histories

[1] The information lifecycle includes creation, collection, use, processing, dissemination, storage, maintenance, disclosure, or disposal (collectively referred to as "processing") of data that may impact privacy.
[2] Definitions for predictability, manageability, and disassociability, which are privacy engineering objectives, can be found in the NIST Privacy Framework at https://www.nist.gov/privacy-framework.

to genetic markers. This data is invaluable for refining algorithms and providing accurate health insights. It also increases cyber risks, including attracting cyber threat actors, and poses unique privacy challenges due to the rapid and expanded data processing capabilities that AI technologies can provide.

Consider a scenario where an AI application is designed to predict disease susceptibility based on an individual's genetic information. Protecting this information from cyber threats is paramount; unauthorized access or manipulation can result in inaccurate predictions, leading to potential mistreatment or misdiagnosis.

Even with secure systems, privacy problems may arise. For example, appropriation of data for non-medical purposes for which an individual did not consent is a genuine concern. While patients might consent to their data being used for one specific purpose, they may not want their data used for other purposes, particularly non-medical ones.

Addressing privacy risk management and cybersecurity risk management together can help organizations govern and manage data appropriately. The CHAI Profile focuses on two main areas:

1. Privacy Risk Management in Healthcare AI: Addresses implications of privacy-related problems individuals may experience as a result of AI-related data processing, as well as the use of AI itself. This includes areas like transparent data usage policies, the nuances of informed consent in an AI context, strategies to ensure data disassociability, and the ethics of workplace surveillance.
2. Cybersecurity Risk Management in Healthcare AI: Addresses specific threats and vulnerabilities that AI systems face, from data breaches to algorithm tampering. It covers protective measures tailored for AI, like securing training data and ensuring the integrity of AI outputs.

Examples of Privacy and Cybersecurity Challenges and Risks at the Nexus of AI and Healthcare

AI adoption has been slow in healthcare, in part because of the complexity of healthcare data and challenges with data management. Organizations must prioritize effective data management to scale AI solutions. Challenges include data gaps, inherent biases in data, difficulties of obtaining data at scale, and sound data governance processes. Effective data management is at the heart of addressing privacy and security risks with healthcare AI.

Data management must be grounded in an effective data mapping exercise and a strong, clear, and efficient data governance process. It is essential to understand the ways the data are used and interact with the AI tools in question so that legal, security, and privacy risks can be accurately assessed and managed. Some issues include:

- What type of data are being used to train AI algorithms or ML models? Are the data raw data or de-identified?

- When using de-identified data, the assurance strength of the de-identification technique needs to be commensurate with the risk of data exposure through re-identification.[3] As described in NIST Special Publication 800-188, *De-Identifying Government Datasets: Techniques and Governance,* "De-identification is 'a general term for any process of removing the association between a set of identifying data and the data subject'[85]...De-identification is not a single technique but a collection of approaches, algorithms, and tools that can be applied to different kinds of data with differing levels of effectiveness."[4] The scope of de-identification techniques includes the removal of specific identifiers, masking data, perturbation, adding noise, and the generation of synthetic data sets. De-identifying data typically creates a tradeoff between privacy and accuracy that must be managed within the context of the intended data use as well as the impact of reidentification.
- Using raw data to train models can alleviate some of the accuracy concerns raised in the previous bullet, however, research has shown that models can be attacked in such a way as to reconstruct and reveal the underlying data on which the model was trained. The use of privacy-enhancing technologies (PETs) such as various types of privacy-enhancing cryptography and differential privacy can be used to prevent these model attacks. Training models relies on data, but healthcare organizations may be reluctant to share or aggregate raw data due to the sensitivity of health information. Techniques such as federated learning, where models are trained locally on the raw data, and only the model updates are aggregated to create a global model for every participant organization's use, can address these raw data sharing concerns. PETs still need to be used, however, to prevent data reconstruction attacks on both the model updates and the final trained model.[5]

- Is a third-party developer of an AI tool (referred to below as an "AI vendor") involved? If so, organizations should engage in a thorough privacy and security review of the AI vendor in accordance with standard processes and include appropriate contractual protections against data misuse. In addition, organizations should assess potential downstream or secondary uses by the AI vendor of any data, or insights derived therefrom. Risks posed by third parties are always essential to address for an effective data privacy and security risk management program. These risks are accentuated with AI because of the power of AI tools.

---

[3] Organizations must ensure de-identification was done in accordance with applicable laws and regulations. In healthcare, the Health Insurance Portability and Accountability Act (HIPAA) sets forth two methods of appropriate deidentification, safe harbor and expert determination. One requirement under the "safe harbor" method is that "the covered entity does not have actual knowledge that the information could be used alone or in combination with other information to identify an individual who is a subject of the information." As a practical matter, this requirement may be more difficult to meet because of the sheer power of AI tools (technology has outpaced regulatory frameworks). In an AI-rich environment, this may drive healthcare providers toward expert determination so that the specific risks of reidentification can be assessed.

[4] https://doi.org/10.6028/NIST.SP.800-188

[5] For more information on privacy-preserving federated learning, see https://www.nist.gov/itl/applied-cybersecurity/privacy-engineering/collaboration-space/blog-series/privacy-preserving.

- Have risks of data tracking tools (cookies, scraping, analytics, etc.) or workplace safety or productivity tools been considered?
    - Tracking tools embedded within a provider's or AI vendor's environment, which may not always be known or discovered absent specific due diligence, pose risk, and these risks are enhanced when combined with the power of AI tools. AI vendors, many of whom may not be subject to the same legal requirements as healthcare entities, have experienced major data breaches because of tracking technologies. Every use of data in an AI environment, including indirect risks posed by data tracking tools, must be assessed in a diligence process.
    - The use of AI tools in the workplace to improve safety outcomes or productivity is increasing. Organizations need to consider the benefits of these tools, as well as the risk of creating a surveillance environment that could affect the dignity and morale of employees and even create unsafe or counterproductive situations through evasions or workarounds by employees.
- Are communications about the AI systems and related data processing effectively providing meaningful information to individuals? Are privacy preferences included in algorithmic design objectives and are the outputs evaluated against these preferences?
- Have security controls been considered to prevent poisoning attacks? For example, enlarging and cleaning the training data set.
- Have access controls for AI processes and application been assessed and integrated with current access controls to minimize possible leaks of all or partial information about the model?
- Has adversarial training been used to strengthen the model against evasion attacks?

Profile Development Approach

Developing the CHAI Profile was a collaborative stakeholder-driven process. Privacy and cybersecurity practitioners from multiple CHAI member organizations contributed to the process to ensure that the CHAI Profile aligns privacy and cybersecurity outcomes with healthcare AI priorities. This section describes how CHAI gathered input and garnered consensus from stakeholders to produce the CHAI Profile.

From June through September 2023, the CHAI Privacy and Security Work Group hosted virtual working sessions with healthcare AI privacy and security stakeholders from government, universities, a non-profit think tank, and industry. During the working sessions, the work group achieved the following two objectives: 1) identified AI-related Healthcare Outcomes and Operational Imperatives for the CHAI Profile (the "Healthcare AI Priorities"), and 2) prioritized privacy and cybersecurity outcomes for each Healthcare AI Priority using the Subcategories in the NIST Privacy and Cybersecurity Framework Cores. CHAI used Cybersecurity Framework v1.1 and Privacy Framework v1.0, the latest versions available at the time. NIST provides resources on the CSF 2.0 website to show the relationships between CSF v1.1

and 2.0.[6] Additionally, the NIST National Cybersecurity Center of Excellence (NCCoE) provides resources for developing and maintaining Community Profiles, like the CHAI Profile.[7]

The CHAI Privacy and Security Work Group Leads invited participants in the CHAI Fall Convening, held in November 2023, to provide additional inputs regarding the priority of outcomes using the Category level of the NIST Privacy (v1.0) and Cybersecurity (v1.1) Frameworks in an effort to ensure highly prioritized outcomes are truly among the highest priorities.

The results from the June-September 2023 works sessions and the November 2024 Fall Convening were synthesized in the CHAI Profile. The resulting 9 Healthcare Outcomes and 3 Operational Imperatives are described in the “Summary of Healthcare Outcomes and Operational Imperatives” section of the CHAI Profile. The prioritized Subcategories are provided in the "Healthcare AI Privacy and Cybersecurity Framework Profile" section of the CHAI Profile. These considerations also informed the privacy and cybersecurity content in the CHAI Responsible AI Guide and Checklist.

Overview of the NIST Frameworks

Each version of the NIST Privacy (v1.0) and Cybersecurity (v1.1) Frameworks (PF & CSF) were created through collaboration between industry and government Both frameworks provide flexible, risk-based, and voluntary guidance based on existing standards, guidelines, and practices to help organizations better understand, manage, reduce, and communicate about privacy and cybersecurity risks. The PF and CSF enable organizations—regardless of size, degree of privacy and cybersecurity risk, or privacy and cybersecurity sophistication—to apply the principles and best practices of risk management to improving privacy, cybersecurity, and resilience. These two frameworks provide a common language for expressing privacy and cybersecurity risk and for conducting management-level privacy and cybersecurity communications among internal and external stakeholders and across an organization, regardless of privacy or cybersecurity expertise.

The PF v1.0 and CSF v1.1 consist of three main components[8]:

1. The Core is a catalog of desired privacy or cybersecurity outcomes. These outcomes are expressed using plain language that is easy to understand regardless of the reader’s role and level of exposure to privacy and cybersecurity concepts. Organizations determine the specific actions they will take to achieve an outcome. The Core complements existing privacy, cybersecurity, and risk management processes and guides organizations in managing and reducing their privacy and cybersecurity risks.
2. Profiles are used to understand, tailor, assess, prioritize, and communicate the Core’s outcomes for organizations and communities. Profiles provide a customized alignment of requirements, objectives, risk appetite, and resources against the desired outcomes of the PF and CSF Cores.

---

[6] The NIST CSF site is available at: https://www.nist.gov/cyberframework. Information and links to mappings and other references are available on the NIST CSF site at: https://www.nist.gov/informative-references. Mappings between CSF versions as well as the CSF and Privacy Framework are available through the National Online Informative References (OLIR) program at: https://csrc.nist.gov/projects/olir/informative-reference-catalog#/.

[7] The NCCoE Framework Resource Center site is located at: https://www.nccoe.nist.gov/framework-resource-center.

[8] The terms Core, Implementation Tiers, Profile, Healthcare AI Priorities, Function, Category, and Subcategory are capitalized when they are used to describe elements of the Cybersecurity Framework throughout this document.

They can be used to identify and prioritize opportunities for improving privacy and cybersecurity in a specific context (e.g., an organization's mission needs or a sector use case like CHAI).

3. Tiers characterize the rigor of an organization's privacy and cybersecurity risk governance and management practices, and they provide context for how an organization views privacy and cybersecurity risk management. Tiers help set the overall tone for how an organization will manage its privacy and cybersecurity risks and understand the extent to which privacy and cybersecurity risk management practices are integrated with broader organizational risk management decisions. (The CHAI Profile focuses on the Core and Profiles.)

The Framework Cores

The PF v1.0 and CSF v1.1 articulate privacy and cybersecurity outcomes using common language that all levels of an organization, from the board and executive level to the individuals with operational roles, can understand. At the top level, the Framework Cores are organized by concurrent and continuous Functions. When considered together, these Functions provide a high-level, strategic view of the lifecycle of an organization's management of privacy and cybersecurity risk. The Functions further subdivide into Categories and Subcategories to convey outcomes for each Function. Tables 1 and 2 present the Functions and Categories in the PF and CSF.

The Core in each framework is also supported by Informative References, which are mappings that indicate relationships between the Core and various standards, guidelines, regulations, and other content to help organizations achieve those outcomes. Informative References can help inform how organizations achieve the Outcomes in the Core. CHAI may wish to create additional Informative Reference unique to this context in the future.

Table 1. Privacy Framework v1.0 Functions and Categories[9]

PF v1.0 Functions	Categories
Identify-P	Inventory and Mapping (ID.IM-P) Business Environment (ID.BE-P) Risk Assessment (ID.RA-P) Data Processing Ecosystem Risk Management (ID.DE-P)
Govern-P	Governance Policies, Processes, and Procedures (GV.PO-P) Risk Management Strategy (GV.RM-P) Awareness and Training (GV.AT-P) Monitoring and Review (GV.MT-P)
Control-P	Data Processing Policies, Processes, and Procedures (CT.PO-P) Data Processing Management (CT.DM-P) Disassociated Processing (CT.DP-P)
Communicate-P	Communication Policies, Processes, and Procedures (CM.PO-P) Data Processing Awareness (CM.AW-P)
Protect-P	Data Protection Policies, Processes, and Procedures (PR.PO-P) Identity Management, Authentication, and Access Control (PR.AC-P) Data Security (PR.DS-P) Maintenance (PR.MA-P) Protective Technology (PR.PT-P)

Table 2. Cybersecurity Framework v1.1 Functions and Categories.

CSF v1.1 Functions	Categories
Identify	Asset Management (ID.AM) Business Environment (ID.BE) Governance (ID.GV) Risk Assessment (ID.RA) Risk Management Strategy (ID.RM) Supply Chain Risk Management (ID.SC)
Protect	Access Control (PR.AC) Awareness and Training (PR.AT) Data Security (PR.DS) Information Protection Processes and Procedures (PR.IP) Maintenance (PR.MA) Protective Technology (PR.PT)
Detect	Anomalies and Events (DE.AE) Security Continuous Monitoring (DE.CM) Detection Processes (DE.DP)
Respond	Response Planning (RS.RP) Communications (RS.CO) Analysis (RS.AN) Mitigation (RS.MI) Improvements (RS.IM)
Recover	Recovery Planning (RC.RP) Improvements (RC.IM) Communications (RC.CO)

[9] The NIST PF v1.0 also points to Detect, Respond, and Recover in the NIST CSF v1.1.

The Categories are further broken down into Subcategories of specific technical or management activities and outcomes. The Healthcare AI Privacy and Cybersecurity Framework Profile section presents the PF and CSF Profile and prioritizes all Subcategories in both frameworks for each Healthcare AI Priority.

Framework Profiles

Framework Profiles help organizations and communities align the Functions, Categories, and Subcategories of the Framework Cores with the business requirements, risk tolerance, and resources. The CHAI Profile offers a prioritization of NIST PF and CSF Subcategories based on priority mission and operational considerations for organizations that use or plan to use AI in the healthcare community. The CHAI Profile serves as a useful starting point for identifying and engaging in discussions about privacy and cybersecurity activities and outcomes that are important to organizations that are using or plan to use AI in the healthcare community. The CHAI Profile offers healthcare organizations the following benefits:

- Describing a shared taxonomy for privacy and cybersecurity risk management and priorities in the context of AI in the healthcare community
- Encouraging common target outcomes that organizations within the healthcare community can use to inform their assessments of privacy and cybersecurity progress when using AI
- Aligning considerations from multiple sources under one framework
- Leveraging expertise across the community
- Minimizing the burden for each organization by providing priorities and outcomes that organizations can use to develop their own Target Profiles

The CHAI Profile is oriented around a set of priorities, which are high-level healthcare outcomes and operational imperatives that enable organizations in the healthcare AI community to succeed. These priorities provide the necessary context for an organization to manage its privacy and cybersecurity risk as it relates to a specific mission need. The CHAI Profile identifies the PF and CSF Subcategories that are especially relevant to each healthcare AI priority and suggests how the PF and CSF Subcategories may be prioritized. An organization can adapt the CHAI Profile priorities and Subcategory prioritization to fit its unique needs.

Summary of Healthcare AI Priorities: Outcomes and Operational Imperatives

The working session discussions resulted in 13 priorities that characterize high-level critical operational needs to an organization to meet its primary reasons for using AI in healthcare (the “Healthcare AI Priorities”). These Healthcare AI Priorities represent community outcomes and operational imperatives. In some cases, the Healthcare AI Priorities are focused on privacy or cybersecurity needs, though the overall set of objectives are broader than privacy or cybersecurity.

The reasons for using AI in healthcare may vary widely organization to organization. These Healthcare AI Priorities appear in alphabetical order under two types, healthcare outcomes and operational imperatives, and are not intended to imply any prioritization. Each Healthcare AI Priority is also aligned to one or more CHAI use cases.

Healthcare Outcomes

These Healthcare AI Priorities represent the reasons healthcare organizations may choose to implement AI capabilities. They focus on how AI can improve patient outcomes and enhance the mission of the organization or the healthcare system overall as well as critical dependencies for implementing AI.

Healthcare AI Priority (Keyword)	Description
Enable Patient Access to Care (Access)	This objective aims to use AI models to get patients in the door by breaking down barriers, increasing convenience, and ensuring that every patient has the opportunity to easily access the right care, at the right time. Examples of how this can be achieved include: implementing virtual care and telemedicine solutions to reach remote or underserved populations, AI- powered appointment scheduling, mobile health applications, AI chatbots and triage systems for personalized guidance, employing predictive analytics for resource planning, developing platforms for health education and empowerment, generating personalized treatment recommendations, and ensuring accessibility and inclusivity for diverse patient populations.
Expedite research initiatives (Research)	This results of AI-powered analyses can promote continuous advancements in healthcare diagnoses and treatment. Organizations encourage collaboration among researchers, clinicians, data scientists, and systems engineers to explore new connections in healthcare AI and examine effective treatments.
Facilitate continuous learning and improvement of the healthcare system (Improvement)	Use of AI can facilitate continuous learning and evaluation of the health care system. Organizations can for example, use AI to gather feedback from healthcare professionals and patients to drive operational and administrative improvements and enhance system effectiveness.
Improve patient outcomes (Outcomes)	This objective aims to enhance patient outcomes, reduce errors, and optimize healthcare delivery by harnessing the potential of AI in improving diagnostic accuracy and treatment efficacy. The aim is to develop AI-driven solutions that can analyze vast amounts of patient data, identify patterns, and provide insights that have the potential to support clinical decision-making. By healthcare AI, the organization can ideally deliver personalized and evidence-based care, improve diagnostic accuracy, and optimize treatment plans which ultimately could lead to improved patient outcomes and enhances quality of care.
Increase efficiency and quality of diagnostic	This objective is to utilize AI to help streamline the diagnostic process in healthcare, by reducing delays and improving resource utilization. AI can also be utilized to improve diagnostic accuracy with improvements in algorithms.

processes (Diagnostics)	By integrating AI algorithms for decision support, triage, and resource optimization, healthcare organizations aim to expedite diagnoses, enhance patient outcomes, and improve the efficiency of diagnostic workflows.
Optimize clinical and administrative resource allocation and efficiency to increase patient value (Optimize)	This objective is to utilize AI to help ensure the effective utilization of healthcare resources to meet patient needs and operational demands. This objective uses AI algorithms to analyze data including patient demographics to inform resource allocation decisions. This can help organizations minimize waste, effectively allocate resources (e.g., right-size costs where expenditures make sense and reducing or eliminating unnecessary or unjustifiable costs), and improve the overall efficiency of healthcare delivery. It could help lead to a better patient experience and outcome, while also lowering costs.
Promote patient engagement and empowerment [Patient-centric approach] (Engagement)	This objective aims to utilize AI strategies and technologies that encourage patient participation, collaboration, and self-management. This may involve developing user-friendly interfaces and patient-centric applications that allow patients to access their data, view personalized insights, and actively engage in their plans all using easily understandable language to convey their health status and enable them to make more informed decisions regarding their care. These applications may also enable remote access and analysis of clinically relevant information from individuals through the use of mobile devices and wearables resulting in a sense of power of an individual over their health. This includes providing educational resources and support materials to enhance health literacy and enable patients to understand the information generated by AI algorithms. An important aspect of patient engagement and empowerment includes mechanisms for patient advocates to provide feedback regarding patient concerns about how their information is processed.
Advance health outcomes for all and reduce disparities (Disparities)	The mission objective highlights reducing disparities and enhancing fairness in healthcare AI. By using AI tools, this mission aims to address biases and provide sensitive, patient-centric care. Transparent AI systems are prioritized to ensure accountability and foster trust. This objective guides all efforts to achieve balanced treatment, access, and outcomes for all patients, ultimately creating a more equal and inclusive healthcare system.
Support Ethical Decision-Making (Ethics)	These objective addresses aligning AI use with the ethical principles in healthcare. It involves organizations ensuring that the AI it uses is responsible, trustworthy, and unbiased. By aiming to use AI aligned with ethical principles in healthcare, then AI technologies themselves can influence clinicians to make decisions in an ethical way. In addition, healthcare AI models are developed to provide insights into the decision-making process, allowing healthcare professionals and patients/caregivers to understand and evaluate the ethical implications of AI-driven recommendations.

Operational Imperatives

These Healthcare AI Priorities represent critical dependencies for the successful use of AI in healthcare organizations. Whereas the Healthcare Outcomes focus on the "why" of AI, these are the things organizations must get right if they want to use AI.

Healthcare AI Priority (Keyword)	Description
Establish and increase transparency about data uses and disclosures in health care (Transparency)	The organization aims to develop AI systems that provide clear explanations of their decision-making process (cognitively). This includes using responsible AI methods such as interpretable algorithms, providing understandable outputs, and maintaining comprehensive documentations. By ensuring transparency and explainable AI, the organization can provide patient and clinician trust, facilitate better understanding of AI-generated insights, and enable effective collaboration between healthcare professionals and AI systems.
Establish Secure Processing Environment (Secure)	Given the sensitivity of healthcare data, robust security and privacy measurements are critical in AI implementation. Organizations aim to establish secure network infrastructure, conduct regular risk assessments, develop incident response mechanisms, implement strong access controls and encryption, adopt proactive threat detection, as part of standard cybersecurity practices. By prioritizing security and privacy, the organization can protect patient privacy, safeguard patient data, prevent unauthorized access, and ensure integrity and availability of AI systems.
Facilitate and maintain compliance with laws, regulations, and standards (Legal)	This objective ensures that the organization adheres to relevant laws and regulations (e.g., HIPAA) governing the use of AI in healthcare. It involves establishing robust frameworks, processes, and practices to meet legal and regulatory requirements, promote ethical AI practices, and maintain high level of privacy and security.

Contents and Use of the CHAI Profile

About the CHAI Profile Contents

The CHAI Profile provides a table that aligns the Healthcare AI Priorities with Subcategories in the PF v1.0 and CSF v1.1. Each Subcategory was assigned High, Moderate, or Other priority, along with a rationale to help an organization understand the relative importance of a Subcategory to a Healthcare AI Priority. Where applicable, Subcategories that share similar outcomes in the PF and CSF are prioritized together. For example, in the Identify (ID) Function of both frameworks, the Business Environment (BE) Category includes a Subcategory about an organization understanding its external roles:

Privacy Framework	Cybersecurity Framework
ID.BE-P1: The organization's role(s) in the data processing ecosystem are identified and communicated.	ID.BE-1: The organization's role in the supply chain is identified and communicated

The CHAI Profile indicates the priority of each Subcategory for each Healthcare Outcome and Operational Imperative using the following designations:

- Three dots (●●●) for High Priority: These represent the most critical Subcategories for enabling a Healthcare AI Priority that should be addressed most immediately given available resources.
- Two dots (●●) for Moderate Priority: These Subcategories should be the next priority after implementing High Priority Subcategories and may become higher priority in certain contexts or environments to implement a given Healthcare AI Priority.
- One dot (●) for Other Implemented Subcategories: Subcategories that are important to the overall cybersecurity of a Healthcare AI Priority but may not require the same level of urgency as higher priority Subcategories. Note that "Other" does not equate to low priority. All Subcategories should receive consideration.

Although organizations should develop privacy and cybersecurity strategies that address all Subcategories, the prioritization provides adaptable guidance that suggests privacy and cybersecurity capabilities that will provide the greatest impact toward meeting Healthcare AI Priorities for organizations in the healthcare community that are using AI. Organizations may further tailor these priorities as needed to address the specific risks in their environment.

How to Use the CHAI Profile

Healthcare organizations can use the CHAI Profile guidance to examine and potentially improve their existing privacy and cybersecurity practices and activities. Examples of how organizations can use the CHAI Profile include:

- Inform executive leadership of CHAI's privacy and cybersecurity expectations and goals
- Align business and operational practices with supporting privacy and cybersecurity activities that have been vetted by CHAI privacy and cybersecurity colleagues
- Benchmark against CHAI expectations when developing the organization's Organizational Current Profile
- Inform the organization's Target Profile(s) or use it as the organization's Target Profile for using AI in healthcare activities
- Facilitate decision making when allocating budget, staffing, and other resources
- Communicate privacy and cybersecurity posture in a consistent way with community partners (e.g., vendors, supply chain, service providers), standards bodies, or regulators

Healthcare organizations can use the CHAI Profile to inform their Organizational Profiles (Current and Target).[10] Organizations that wish to use the CHAI Profile for their Organizational Target Profile can take the following steps:

1. Examine how the organizations priorities for using AI align with the 13 Healthcare AI Priorities in the CHAI Profile. Organizations may apply the priorities in the CHAI Profile, adapt them to better fit their organizational needs, or develop their own.
2. Prioritize Healthcare AI Priorities based on their requirements and strategic goals. Regardless of whether the organization uses the Healthcare AI Priorities provided in the CHAI Profile or a version of their own, prioritization helps with later use for activities such as strategic planning.
3. Identify applicable Informative References. Healthcare organizations should consider any constraints or guidance (e.g., applicable state laws, policies, standards), risks, and other influencing factors that inform how it achieves the outcomes in each Subcategory. Informative References may also include the organization's own policies, standards, procedures, and guidance.
4. Refine Subcategory prioritizations as needed. Organizations may need to adjust the priority level of Subcategories in the CHAI Profile to align with the risks in their operating environment.
5. Determine additional implementation guidance as needed. Organizations can document rationale, considerations, and any additional information their organization finds useful for using their Target Profile.

Organizations can use their own Current and Target Profiles together identify any gaps between their current privacy and cybersecurity capabilities and their target state. This gap analysis can help an organization determine if re-allocation of privacy and cybersecurity resources toward higher priority capabilities would help them achieve those prioritized Subcategories and therefore, better achieve their organization's Healthcare AI Priorities.

[10] An Organizational Profile describes an organization's current and/or target cybersecurity posture in terms of cybersecurity outcomes in the CSF Core. Community Profiles, such as the CHAI Profile (this document) can help organization created Organizational Profiles. The NIST Quick-Start Guide for Creating and Using Organizational Profiles, which is available here: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1301.pdf.

The CHAI Profile

Framework Subcategories		Access	Diagnostics	Disparities	Engagement	Ethics	Improvement	Legal	Outcomes	Optimize	Research	Secure	Transparency	Mapped to One or More Consider-ations?
Privacy Framework	Cybersecurity Framework													
	ID.AM-1: Physical devices and systems within the organization are inventoried	●●	●●●	●●	●●	●●	●●	●●●	●●	●●●	●●	●●●	●●	YES
ID.IM-P1: Systems / products / services that process data are inventoried.		●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●	●●	●●●	●●●	●●	YES
	ID.AM-2: Software platforms and applications within the organization are inventoried	●●●	●●●	●●	●●	●●	●●●	●●●	●●	●●●	●●●	●●●	●●	YES
ID.IM-P2: Owners or operators (e.g., the organization or third parties such as service providers, partners, customers, and developers) and their roles with respect to the		●●●	●●●	●●●	●●●	●●	●●●	●●	●●●	●●●	●●●	●●●	●●●	YES

Framework Subcategories		Access	Diagnostics	Disparities	Engagement	Ethics	Improvement	Legal	Outcomes	Optimize	Research	Secure	Transparency	Mapped to One or More Consider-ations?
Privacy Framework	Cybersecurity Framework													
systems/products/services and components (e.g., internal or external) that process data are inventoried.														
	ID.AM-3: Organizational communication and data flows are mapped	●●●	●●●	●●	●●	●●	●●●	●●	●●	●●●	●●●	●●●	●●	YES
ID.IM-P3: Categories of individuals (e.g., customers, employees or prospective employees, consumers) whose data are being processed are inventoried		●●	●●	●●	●●	●●	●	●	●●	●●	●●	●●	●●	
	ID.AM-4: External information systems are catalogued	●●	●●	●	●	●	●●	●	●	●●	●●	●●	●	YES

Framework Subcategories		Access	Diagnostics	Disparities	Engagement	Ethics	Improvement	Legal	Outcomes	Optimize	Research	Secure	Transparency	Mapped to One or More Consider-ations?
Privacy Framework	Cybersecurity Framework													
ID.IM-P4: Data actions of the systems / products / services are inventoried.		●●●	●●●	●●●	●●●	●●●	●●	●●	●●	●	●●	●●●	●●	YES
	ID.AM-5: Resources (e.g., hardware, devices, data, time, personnel, and software) are prioritized based on their classification, criticality, and business value	●●	●●	●	●	●	●	●	●	●●	●	●●	●	
ID.IM-P5: The purposes for the data actions are inventoried.		●●●	●●●	●●●	●●●	●●●	●●	●●	●●●	●●	●●	●●●	●●	YES
ID.IM-P6: Data elements within the data actions are inventoried.		●●●	●●●	●●●	●●	●●	●●	●●	●●●	●●	●●●	●●●	●●	YES
ID.IM-P7: The data processing environment is identified (e.g., geographic location,		●●●	●●●	●●	●●	●●	●●	●●	●●	●●	●●	●●●	●●	YES

Framework Subcategories		Access	Diagnostics	Disparities	Engagement	Ethics	Improvement	Legal	Outcomes	Optimize	Research	Secure	Transparency	Mapped to One or More Consider-ations?
Privacy Framework	Cybersecurity Framework													
internal, cloud, third parties).														
ID.IM-P8: Data processing is mapped, illustrating the data actions and associated data elements for systems/products/services, including components; roles of the component owners/operators; and interactions of individuals or third parties with the systems/products/services.		●●●	●●●	●●●	●●	●●	●●	●●	●●●	●●●	●●	●●●	●●●	YES
ID.BE-P1: The organization's role(s) in the data processing ecosystem are identified and communicated.	ID.BE-1: The organization's role in the supply chain is identified and communicated	●●	●●	●●	●	●	●	●●	●	●	●	●	●	YES
	ID.BE-2: The organization's place in critical infrastructure	●●	●●	●●	●	●	●	●●	●	●	●	●	●	

Framework Subcategories		Access	Diagnostics	Disparities	Engagement	Ethics	Improvement	Legal	Outcomes	Optimize	Research	Secure	Transparency	Mapped to One or More Consider-ations?
Privacy Framework	Cybersecurity Framework													
	and its industry sector is identified and communicated													
ID.BE-P2: Priorities for organizational mission, objectives, and activities are established and communicated.	ID.BE-3: Priorities for organizational mission, objectives, and activities are established and communicated	●●	●●	●●	●●	●●	●●	●	●●	●●	●	●●	●●	YES
ID.BE-P3: Systems / products / services that support organizational priorities are identified and key requirements communicated.		●	●	●	●	●	●	●	●	●	●	●	●	
	ID.BE-4: Dependencies and critical functions for delivery of critical services are established	●●	●●●	●●	●●	●●	●●●	●●	●●●	●●	●●●	●●●	●●	YES
	ID.BE-5: Resilience requirements to support delivery of critical services are established for all	●	●●	●●	●●	●●	●●	●●	●●●	●	●●	●●	●●	YES

Framework Subcategories		Access	Diagnostics	Disparities	Engagement	Ethics	Improvement	Legal	Outcomes	Optimize	Research	Secure	Transparency	Mapped to One or More Consider-ations?
Privacy Framework	Cybersecurity Framework													
	operating states (e.g. under duress/attack, during recovery, normal operations)													
	ID.RA-1: Asset vulnerabilities are identified and documented	●●	●●	●●	●●	●●	●●	●●●	●●	●	●●	●●	●	YES
ID.RA-P1: Contextual factors related to the systems/products/services and the data actions are identified (e.g., individuals' demographics and privacy interests or perceptions, data sensitivity and/or types, visibility of data processing to individuals and third parties).		●●●	●●●	●●●	●●●	●●●	●	●●	●●●	●●●	●●	●●●	●●●	YES
	ID.RA-2: Cyber threat intelligence is received from information	●	●	●	●	●	●	●	●	●	●	●●	●	

Framework Subcategories		Access	Diagnostics	Disparities	Engagement	Ethics	Improvement	Legal	Outcomes	Optimize	Research	Secure	Transparency	Mapped to One or More Consider-ations?
Privacy Framework	Cybersecurity Framework													
	sharing forums and sources													
ID.RA-P2: Data analytic inputs and outputs are identified and evaluated for bias.		●●●	●●●	●●●	●●●	●●●	●●	●●	●●	●●	●●	●●	●●	YES
	ID.RA-3: Threats, both internal and external, are identified and documented	●●●	●●●	●●	●●	●●	●●●	●●	●●	●●	●●●	●●●	●●	YES
ID.RA-P3: Potential problematic data actions and associated problems are identified.		●●●	●●●	●●	●	●●●	●	●●	●●	●●	●	●●	●	YES
ID.RA-P4: Problematic data actions, likelihoods, and impacts are used to determine and prioritize risk.	ID.RA-4 | ID.RA-5: Potential business impacts and likelihoods are identified | Threats, vulnerabilities, likelihoods, and impacts are used to determine risk	●●●	●●●	●	●	●●●	●	●●	●●	●●	●	●●	●	
		●●	●●	●●	●●	●●	●●●	●●	●●	●●	●●●	●●	●●	YES

Framework Subcategories		Access	Diagnostics	Disparities	Engagement	Ethics	Improvement	Legal	Outcomes	Optimize	Research	Secure	Transparency	Mapped to One or More Consider-ations?
Privacy Framework	Cybersecurity Framework													
ID.RA-P5: Risk responses are identified, prioritized, and implemented.	ID.RA-6: Risk responses are identified and prioritized	●●●	●●●	●	●●	●●●	●●	●●	●●	●	●	●●	●	YES
ID.DE-P1: Data processing ecosystem risk management policies, processes, and procedures are identified, established, assessed, managed, and agreed to by organizational stakeholders.	ID.SC-1: Cyber supply chain risk management processes are identified, established, assessed, managed, and agreed to by organizational stakeholders	●●	●●	●●	●●	●●	●	●●●	●	●	●●	●●	●	YES
ID.DE-P2: Data processing ecosystem parties (e.g., service providers, customers, partners, product manufacturers, application developers) are identified, prioritized, and assessed using a privacy risk assessment process.	ID.SC-2: Suppliers and third party partners of information systems, components, and services are identified, prioritized, and assessed using a cyber supply chain risk assessment process	●●●	●●●	●●	●●	●●	●●	●●	●●	●●	●	●●●	●●	YES

Framework Subcategories		Access	Diagnostics	Disparities	Engagement	Ethics	Improvement	Legal	Outcomes	Optimize	Research	Secure	Transparency	Mapped to One or More Consider-ations?
Privacy Framework	Cybersecurity Framework													
ID.DE-P3: Contracts with data processing ecosystem parties are used to implement appropriate measures designed to meet the objectives of an organization's privacy program.	ID.SC-3: Contracts with suppliers and third-party partners are used to implement appropriate measures designed to meet the objectives of an organization's cybersecurity program and Cyber Supply Chain Risk Management Plan.	●●●	●●●	●●	●●	●●	●●	●●●	●●	●●	●●	●●●	●●	YES
ID.DE-P4: Interoperability frameworks or similar multi-party approaches are used to manage data processing ecosystem privacy risks.		●	●	●●	●●	●	●	●	●●	●●	●	●●	●●	
ID.DE-P5: Data processing ecosystem parties are routinely assessed using audits, test results, or forms of evaluations to confirm they are meeting their	ID.SC-4: Suppliers and third-party partners are routinely assessed using audits, test results, or forms of evaluations to confirm	●●	●●●	●●	●●	●●	●●	●●	●●	●●	●	●●●	●●	YES

Framework Subcategories		Access	Diagnostics	Disparities	Engagement	Ethics	Improvement	Legal	Outcomes	Optimize	Research	Secure	Transparency	Mapped to One or More Consider-ations?
Privacy Framework	Cybersecurity Framework													
contractual, interoperability framework, or obligations.	they are meeting their contractual obligations.													
	ID.SC-5: Response and recovery planning and testing are conducted with suppliers and third-party providers	●●	●●	●●	●●	●●	●	●●	●●	●●	●	●●	●●	YES
GV.PO-P1: Organizational privacy values and policies (e.g., conditions on data processing such as data uses or retention periods, individuals' prerogatives with respect to data processing) are established and communicated.	ID.GV-1: Organizational cybersecurity policy is established and communicated	●●●	●●●	●●●	●●●	●●	●●	●●●	●●●	●●●	●●	●●	●●●	YES

Framework Subcategories		Access	Diagnostics	Disparities	Engagement	Ethics	Improvement	Legal	Outcomes	Optimize	Research	Secure	Transparency	Mapped to One or More Consider-ations?
Privacy Framework	Cybersecurity Framework													
GV.PO-P2: Processes to instill organizational privacy values within system/product/service development and operations are established and in place.		●●●	●●●	●●●	●●●	●●	●	●	●●●	●●●	●	●	●●	YES
GV.PO-P3: Roles and responsibilities for the workforce are established with respect to privacy.	ID.AM-6: Cybersecurity roles and responsibilities for the entire workforce and third-party stakeholders (e.g., suppliers, customers, partners) are established	●●●	●●●	●	●	●	●	●	●●	●	●	●●	●	YES
GV.PO-P4: Privacy roles and responsibilities are coordinated and aligned with third-party stakeholders (e.g., service	ID.GV-2: Cybersecurity roles and responsibilities are coordinated and aligned with internal roles and external partners	●●	●●	●	●●	●	●	●●	●	●	●	●●	●	YES

Framework Subcategories		Access	Diagnostics	Disparities	Engagement	Ethics	Improvement	Legal	Outcomes	Optimize	Research	Secure	Transparency	Mapped to One or More Consider-ations?
Privacy Framework	Cybersecurity Framework													
providers, customers, partners).														
GV.PO-P5: Legal, regulatory, and contractual requirements regarding privacy are understood and managed.	ID.GV-3: Legal and regulatory requirements regarding cybersecurity, including privacy and civil liberties obligations, are understood and managed	●●●	●●●	●●●	●●●	●●	●●	●●●	●●●	●●●	●●	●●●	●●●	YES
GV.PO-P6: Governance and risk management policies, processes, and procedures address privacy risks.	ID.GV-4: Governance and risk management processes address cybersecurity risks	●●●	●●●	●●	●●	●●	●●	●●●	●●	●	●●	●●●	●●●	YES
GV.RM-P1: Risk management processes are established, managed, and agreed to by organizational stakeholders.	ID.RM-1: Risk management processes are established, managed, and agreed to by organizational stakeholders	●●●	●●●	●	●●	●	●●	●●	●●	●●	●●	●●	●●	YES

Framework Subcategories		Access	Diagnostics	Disparities	Engagement	Ethics	Improvement	Legal	Outcomes	Optimize	Research	Secure	Transparency	Mapped to One or More Consider-ations?
Privacy Framework	Cybersecurity Framework													
GV.RM-P2: Organizational risk tolerance is determined and clearly expressed.	ID.RM-2: Organizational risk tolerance is determined and clearly expressed	●●	●●	●	●●	●	●●	●●	●●	●●	●●	●●	●●	YES
GV.RM-P3: The organization's determination of risk tolerance is informed by its role(s) in the data processing ecosystem	ID.RM-3: The organization's determination of risk tolerance is informed by its role in critical infrastructure and sector specific risk analysis	●●	●	●	●●	●	●	●●	●●	●	●	●	●●	YES
GV.AT-P1: The workforce is informed and trained on its roles and responsibilities.	PR.AT-1: All users are informed and trained	●●	●●	●●	●●	●●	●	●●	●●	●●	●●	●●	●●	YES
	PR.AT-2: Privileged users understand their roles and responsibilities	●●	●●	●●	●●	●●	●●	●●	●●	●●	●●	●●	●●	
GV.AT-P2: Senior executives understand	PR.AT-4: Senior executives understand	●●	●●	●●	●●	●●	●●	●●	●●	●●	●●	●●	●●	

Framework Subcategories		Access	Diagnostics	Disparities	Engagement	Ethics	Improvement	Legal	Outcomes	Optimize	Research	Secure	Transparency	Mapped to One or More Consider-ations?
Privacy Framework	Cybersecurity Framework													
their roles and responsibilities.	their roles and responsibilities													
GV.AT-P3: Privacy personnel understand their roles and responsibilities.	PR.AT-5: Physical and cybersecurity personnel understand their roles and responsibilities	•	••	••	••	••	••	••	••	•	••	•••	••	YES
GV.AT-P4: Third parties (e.g., service providers, customers, partners) understand their roles and responsibilities.	PR.AT-3: Third-party stakeholders (e.g., suppliers, customers, partners) understand their roles and responsibilities	•	•	••	••	•	•	••	••	••	•	•••	••	YES
GV.MT-P1: Privacy risk is re-evaluated on an ongoing basis and as key factors, including the organization's business environment (e.g., introduction of new technologies), governance (e.g., legal obligations, risk tolerance), data processing, and		••	••	••	••	••	•	••	••	•	••	••	•	YES

Framework Subcategories		Access	Diagnostics	Disparities	Engagement	Ethics	Improvement	Legal	Outcomes	Optimize	Research	Secure	Transparency	Mapped to One or More Consider-ations?
Privacy Framework	Cybersecurity Framework													
systems/products/services change.														
GV.MT-P2: Privacy values, policies, and training are reviewed and any updates are communicated.		•	•	•	•	•	•	••	•	•	•	•	•	YES
GV.MT-P3: Policies, processes, and procedures for assessing compliance with legal requirements and privacy policies are established and in place.		•	•	•	•	•	•	••	•	•	•	•	•	YES

Framework Subcategories		Access	Diagnostics	Disparities	Engagement	Ethics	Improvement	Legal	Outcomes	Optimize	Research	Secure	Transparency	Mapped to One or More Consider-ations?
Privacy Framework	Cybersecurity Framework													
GV.MT-P4: Policies, processes, and procedures for communicating progress on managing privacy risks are established and in place.		●●	●●	●	●	●	●	●●	●	●●	●●	●	●	YES
GV.MT-P5: Policies, processes, and procedures are established and in place to receive, analyze, and respond to problematic data actions disclosed to the organization from internal and external sources (e.g., internal discovery, privacy researchers, professional events).		●●	●●	●	●	●	●	●●●	●●	●	●	●●	●●	YES
GV.MT-P6: Policies, processes, and procedures incorporate lessons learned from		●	●●●	●●	●	●●	●	●●●	●●	●	●	●●	●●●	YES

Framework Subcategories		Access	Diagnostics	Disparities	Engagement	Ethics	Improvement	Legal	Outcomes	Optimize	Research	Secure	Transparency	Mapped to One or More Consider-ations?
Privacy Framework	Cybersecurity Framework													
problematic data actions.														
GV.MT-P7: Policies, processes, and procedures for receiving, tracking, and responding to complaints, concerns, and questions from individuals about organizational privacy practices are established and in place.		●●	●●	●●	●●	●●	●	●●●	●●	●	●	●	●●●	YES
CT.PO-P1: Policies, processes, and procedures for authorizing data processing (e.g., organizational decisions, individual consent), revoking authorizations, and maintaining authorizations are		●●	●●●	●●	●●	●●●	●	●●●	●●	●●	●●	●●●	●●	YES

Framework Subcategories		Access	Diagnostics	Disparities	Engagement	Ethics	Improvement	Legal	Outcomes	Optimize	Research	Secure	Transparency	Mapped to One or More Consider-ations?
Privacy Framework	Cybersecurity Framework													
established and in place.														
CT.PO-P2: Policies, processes, and procedures for enabling data review, transfer, sharing or disclosure, alteration, and deletion are established and in place (e.g., to maintain data quality, manage data retention).		●●●	●●●	●●●	●●●	●●●	•	●●●	●●●	●●	●●	●●●	●●●	YES
CT.PO-P3: Policies, processes, and procedures for enabling individuals' data processing preferences and requests are		●●	●●	●●	●●●	●●●	•	●●●	●●	●●●	●●	●●	●●●	YES

Framework Subcategories		Access	Diagnostics	Disparities	Engagement	Ethics	Improvement	Legal	Outcomes	Optimize	Research	Secure	Transparency	Mapped to One or More Consider-ations?
Privacy Framework	Cybersecurity Framework													
established and in place.														
CT.PO-P4: A data life cycle to manage data is aligned and implemented with the system development life cycle to manage systems.	PR.IP-2: A System Development Life Cycle to manage systems is implemented	●●●	●●●	●●●	●●●	●●	●	●	●●●	●●●	●	●●●	●●●	YES
CT.DM-P1: Data elements can be accessed for review.		●●	●●	●●	●●	●●	●	●	●●●	●●●	●	●●	●	YES
CT.DM-P2: Data elements can be accessed for transmission or disclosure.		●●	●●	●●	●●	●●	●	●	●●●	●●●	●	●●	●	YES
CT.DM-P3: Data elements can be accessed for alteration.		●●●	●●●	●●●	●●●	●●●	●	●	●●●	●●●	●	●●	●	YES

Framework Subcategories		Access	Diagnostics	Disparities	Engagement	Ethics	Improvement	Legal	Outcomes	Optimize	Research	Secure	Transparency	Mapped to One or More Consider-ations?
Privacy Framework	Cybersecurity Framework													
CT.DM-P4: Data elements can be accessed for deletion.		●●●	●●●	●●●	●●●	●●●	●	●	●●●	●●●	●	●●	●	YES
CT.DM-P5: Data are destroyed according to policy.	PR.IP-6: Data is destroyed according to policy	●●●	●●●	●●●	●●●	●●●	●●	●	●●●	●●●	●	●●●	●	YES
CT.DM-P6: Data are transmitted using standardized formats.		●●	●●	●●	●●	●●●	●	●	●●	●●	●	●●	●	YES
CT.DM-P7: Mechanisms for transmitting processing permissions and related data values with data elements are established and in place.		●●	●	●●	●●	●●	●	●	●●	●●	●	●	●	YES
CT.DM-P8: Audit/log records are determined, documented, implemented, and reviewed in accordance with policy and	PR.PT-1: Audit/log records are determined, documented, implemented, and reviewed in accordance with policy	●●	●●	●●	●●	●●	●	●●	●●	●●	●●	●●	●●	YES

Framework Subcategories		Access	Diagnostics	Disparities	Engagement	Ethics	Improvement	Legal	Outcomes	Optimize	Research	Secure	Transparency	Mapped to One or More Consider-ations?
Privacy Framework	Cybersecurity Framework													
incorporating the principle of data minimization.														
CT.DM-P9: Technical measures implemented to manage data processing are tested and assessed.		•	•	•	•	•	•	••	•	••	•	•	••	YES
CT.DM-P10: Stakeholder privacy preferences are included in algorithmic design objectives and outputs are evaluated against these preferences.		••	•••	••	••	•••	••	•	•••	••	••	•	••	YES
CT.DP-P1: Data are processed to limit observability and linkability (e.g., data actions take place on local devices, privacy-		•	••	•	••	••	••	•	••	•	••	••	•	YES

Framework Subcategories		Access	Diagnostics	Disparities	Engagement	Ethics	Improvement	Legal	Outcomes	Optimize	Research	Secure	Transparency	Mapped to One or More Consider-ations?
Privacy Framework	Cybersecurity Framework													
preserving cryptography).														
CT.DP-P2: Data are processed to limit the identification of individuals (e.g., de-identification privacy techniques, tokenization).		●●●	●●●	●●	●●	●●	●●	●●	●●●	●●	●●	●●	●●	YES
CT.DP-P3: Data are processed to limit the formulation of inferences about individuals' behavior or activities (e.g., data processing is decentralized, distributed architectures).		●●	●●	●●●	●●	●●●	●	●	●●	●●	●	●●	●	YES
CT.DP-P4: System or device configurations permit selective collection or		●●	●●	●●	●●	●	●	●	●●	●●	●	●●●	●	

Framework Subcategories		Access	Diagnostics	Disparities	Engagement	Ethics	Improvement	Legal	Outcomes	Optimize	Research	Secure	Transparency	Mapped to One or More Consider-ations?
Privacy Framework	Cybersecurity Framework													
disclosure of data elements.														
CT.DP-P5: Attribute references are substituted for attribute values.		•	•	•	•	•	•	•	•	•	•	•	•	
CM.PO-P1: Transparency policies, processes, and procedures for communicating data processing purposes, practices, and associated privacy risks are established and in place.		••	••	••	••	•	•	••	••	••	•	•	•••	YES
CM.PO-P2: Roles and responsibilities (e.g., public relations) for communicating data processing purposes, practices, and associated privacy risks are established.		••	••	•	••	•	•	•	•	•	•	••	••	

Framework Subcategories		Access	Diagnostics	Disparities	Engagement	Ethics	Improvement	Legal	Outcomes	Optimize	Research	Secure	Transparency	Mapped to One or More Consider-ations?
Privacy Framework	Cybersecurity Framework													
CM.AW-P1: Mechanisms (e.g., notices, internal or public reports) for communicating data processing purposes, practices, associated privacy risks, and options for enabling individuals' data processing preferences and requests are established and in place.		●●●	●●●	●●	●●●	●●	●	●	●●●	●●	●	●●	●●●	YES
CM.AW-P2: Mechanisms for obtaining feedback from individuals (e.g., surveys or focus groups) about data processing and associated privacy risks are established and in place.		●●	●●	●●	●●●	●	●●	●	●●●	●●	●	●●	●	YES
CM.AW-P3: System / product / service		●●●	●●●	●●●	●●●	●●	●	●	●●●	●●●	●	●	●●●	YES

Framework Subcategories		Access	Diagnostics	Disparities	Engagement	Ethics	Improvement	Legal	Outcomes	Optimize	Research	Secure	Transparency	Mapped to One or More Consider-ations?
Privacy Framework	Cybersecurity Framework													
design enables data processing visibility.														
CM.AW-P4: Records of data disclosures and sharing are maintained and can be accessed for review or transmission/disclosure.		●●	●●●	●●	●●	●	●	●	●●●	●●	●●	●●	●●●	
CM.AW-P5: Data corrections or deletions can be communicated to individuals or organizations (e.g., data sources) in the data processing ecosystem.		●●	●●●	●●●	●●	●●	●	●	●●●	●	●	●●	●●	YES
CM.AW-P6: Data provenance and lineage are maintained and can be accessed for review or		●●●	●●●	●●●	●●●	●●	●●●	●	●●●	●●●	●●●	●●●	●	YES

Framework Subcategories		Access	Diagnostics	Disparities	Engagement	Ethics	Improvement	Legal	Outcomes	Optimize	Research	Secure	Transparency	Mapped to One or More Consider-ations?
Privacy Framework	Cybersecurity Framework													
transmission/disclosure.														
CM.AW-P7: Impacted individuals and organizations are notified about a privacy breach or event.		●●●	●●●	●●	●●	●	●●	●●	●●	●●	●	●●●	●	YES
CM.AW-P8: Individuals are provided with mitigation mechanisms (e.g., credit monitoring, consent withdrawal, data alteration or deletion) to address impacts of problematic data actions.		●●	●●	●●	●●	●	●●	●	●●	●●	●	●●●	●	YES
PR.PO-P1: A baseline configuration of information technology is created and maintained	PR.IP-1: A baseline configuration of information technology/industrial control systems is	●●●	●●●	●	●●	●	●●●	●	●●	●	●●●	●●●	●	YES

Framework Subcategories		Access	Diagnostics	Disparities	Engagement	Ethics	Improvement	Legal	Outcomes	Optimize	Research	Secure	Transparency	Mapped to One or More Consider-ations?
Privacy Framework	Cybersecurity Framework													
incorporating security principles (e.g., concept of least functionality).	created and maintained incorporating security principles (e.g. concept of least functionality)													
PR.PO-P2: Configuration change control processes are established and in place.	PR.IP-3: Configuration change control processes are in place	●●●	●●●	●●	●●	●●	●●	●	●	●	●●	●●●	●	YES
PR.PO-P3: Backups of information are conducted, maintained, and tested.	PR.IP-4: Backups of information are conducted, maintained, and tested	●	●●●	●●	●●	●●	●●	●	●●	●	●●●	●●●	●	
PR.PO-P4: Policy and regulations regarding the physical operating environment for organizational assets are met.	PR.IP-5: Policy and regulations regarding the physical operating environment for organizational assets are met	●	●	●	●	●	●	●●●	●	●	●	●●●	●	
PR.PO-P5: Protection processes are improved.	PR.IP-7: Protection processes are improved	●●●	●●●	●	●	●	●	●	●●	●	●	●●●	●	

Framework Subcategories		Access	Diagnostics	Disparities	Engagement	Ethics	Improvement	Legal	Outcomes	Optimize	Research	Secure	Transparency	Mapped to One or More Consider-ations?
Privacy Framework	Cybersecurity Framework													
PR.PO-P6: Effectiveness of protection technologies is shared.	PR.IP-8: Effectiveness of protection technologies is shared	•	•	•	•	•	•	•	•	•	•	••	•	
PR.PO-P7: Response plans (Incident Response and Business Continuity) and recovery plans (Incident Recovery and Disaster Recovery) are established, in place, and managed.	PR.IP-9: Response plans (Incident Response and Business Continuity) and recovery plans (Incident Recovery and Disaster Recovery) are in place and managed	•••	•••	•	•	•	•	••	••	•	•	•••	•	YES
PR.PO-P8: Response and recovery plans are tested.	PR.IP-10: Response and recovery plans are tested	•••	•••	•	•	•	•	••	••	•	•	•••	•	YES
PR.PO-P9: Privacy procedures are included in human resources practices (e.g., deprovisioning, personnel screening).	PR.IP-11: Cybersecurity is included in human resources practices (e.g., deprovisioning, personnel screening)	•	•	•	•	•	•	•	•	•	••	•••	•	

Framework Subcategories		Access	Diagnostics	Disparities	Engagement	Ethics	Improvement	Legal	Outcomes	Optimize	Research	Secure	Transparency	Mapped to One or More Consider-ations?
Privacy Framework	Cybersecurity Framework													
PR.PO-P10: A vulnerability management plan is developed and implemented.	PR.IP-12: A vulnerability management plan is developed and implemented	●●●	●●●	●●	●●	●	●	●	●●	●●	●	●●●	●●	YES
PR.AC-P1: Identities and credentials are issued, managed, verified, revoked, and audited for authorized individuals, processes, and devices.	PR.AC-1: Identities and credentials are issued, managed, verified, revoked, and audited for authorized devices, users and processes	●●●	●●●	●●●	●●●	●●	●●	●●	●●●	●●●	●●	●●●	●●●	YES
PR.AC-P2: Physical access to data and devices is managed.	PR.AC-2: Physical access to assets is managed and protected	●●●	●●●	●●●	●●	●	●	●	●●●	●●●	●	●●●	●●●	
PR.AC-P3: Remote access is managed.	PR.AC-3: Remote access is managed	●●●	●●●	●●●	●●●	●	●	●	●●●	●●●	●●	●●●	●●●	YES
PR.AC-P4: Access permissions and authorizations are managed, incorporating the principles of least	PR.AC-4: Access permissions and authorizations are managed, incorporating the principles of least	●●●	●●●	●●●	●●●	●●	●●	●●	●●●	●●●	●●	●●●	●●●	YES

Framework Subcategories		Access	Diagnostics	Disparities	Engagement	Ethics	Improvement	Legal	Outcomes	Optimize	Research	Secure	Transparency	Mapped to One or More Consider-ations?
Privacy Framework	Cybersecurity Framework													
privilege and separation of duties.	privilege and separation of duties													
PR.AC-P5: Network integrity is protected (e.g., network segregation, network segmentation).	PR.AC-5: Network integrity is protected (e.g., network segregation, network segmentation)	●●●	●●●	●	●●	●	●	●	●	●	●	●●●	●	YES
PR.AC-P6: Individuals and devices are proofed and bound to credentials, and authenticated commensurate with the risk of the transaction (e.g., individuals' security and privacy risks and organizational risks).	PR.AC-6 | PR.AC-7: Identities are proofed and bound to credentials and asserted in interactions | Users, devices, and assets are authenticated (e.g., single-factor, multi-factor) commensurate with the risk of the transaction (e.g., individuals' security and privacy risks and organizational risks)	●●	●●●	●●	●●	●●	●●	●	●●●	●●	●●	●●●	●●	
PR.DS-P1: Data-at-rest are protected.	PR.DS-1: Data-at-rest is protected	●●●	●●●	●●	●●	●	●●	●●	●●●	●●	●	●●●	●●	YES

Framework Subcategories		Access	Diagnostics	Disparities	Engagement	Ethics	Improvement	Legal	Outcomes	Optimize	Research	Secure	Transparency	Mapped to One or More Consider-ations?
Privacy Framework	Cybersecurity Framework													
PR.DS-P2: Data-in-transit are protected.	PR.DS-2: Data-in-transit is protected	●●●	●●●	●●	●●	●	●●	●●	●●●	●●	●	●●●	●●	YES
PR.DS-P3: Systems/products/services and associated data are formally managed throughout removal, transfers, and disposition.	PR.DS-3: Assets are formally managed throughout removal, transfers, and disposition	●●	●●	●	●	●	●	●	●	●	●●	●●●	●	
PR.DS-P4: Adequate capacity to ensure availability is maintained.	PR.DS-4: Adequate capacity to ensure availability is maintained	●●	●●	●	●	●	●	●	●	●	●	●●●	●	
PR.DS-P5: Protections against data leaks are implemented.	PR.DS-5: Protections against data leaks are implemented	●●●	●●●	●	●	●	●	●●	●●	●	●●●	●●●	●	YES
PR.DS-P6: Integrity checking mechanisms are used to verify software, firmware, and information integrity.	PR.DS-6: Integrity checking mechanisms are used to verify software, firmware, and information integrity	●●●	●●●	●	●●	●	●	●	●	●	●	●●●	●	

Framework Subcategories		Access	Diagnostics	Disparities	Engagement	Ethics	Improvement	Legal	Outcomes	Optimize	Research	Secure	Transparency	Mapped to One or More Consider-ations?
Privacy Framework	Cybersecurity Framework													
PR.DS-P7: The development and testing environment(s) are separate from the production environment.	PR.DS-7: The development and testing environment(s) are separate from the production environment	●●●	●●●	●	●	●	●●	●●	●●	●●	●●	●●●	●	YES
PR.DS-P8: Integrity checking mechanisms are used to verify hardware integrity.	PR.DS-8: Integrity checking mechanisms are used to verify hardware integrity	●●	●●	●	●	●	●	●	●	●	●	●●●	●	
PR.MA-P1: Maintenance and repair of organizational assets are performed and logged, with approved and controlled tools.	PR.MA-1: Maintenance and repair of organizational assets are performed and logged, with approved and controlled tools	●●	●●	●	●	●	●	●	●	●	●●●	●●●	●	
PR.MA-P2: Remote maintenance of organizational assets is approved, logged, and performed in a manner that prevents unauthorized access.	PR.MA-2: Remote maintenance of organizational assets is approved, logged, and performed in a manner that prevents unauthorized access	●●	●●●	●	●●	●	●	●	●●	●	●●	●●●	●	

Framework Subcategories		Access	Diagnostics	Disparities	Engagement	Ethics	Improvement	Legal	Outcomes	Optimize	Research	Secure	Transparency	Mapped to One or More Consider-ations?
Privacy Framework	Cybersecurity Framework													
PR.PT-P1: Removable media is protected and its use restricted according to policy.	PR.PT-2: Removable media is protected and its use restricted according to policy	•	••	•	•	•	•	•	••	•	••	•••	•	YES
PR.PT-P2: The principle of least functionality is incorporated by configuring systems to provide only essential capabilities.	PR.PT-3: The principle of least functionality is incorporated by configuring systems to provide only essential capabilities	•••	•••	•	••	•	•	•	••	••	•	•••	•	YES
PR.PT-P3: Communications and control networks are protected.	PR.PT-4: Communications and control networks are protected	•••	•••	•	•	•	•	•	•	•	•	•••	•	YES
PR.PT-P4: Mechanisms (e.g., failsafe, load balancing, hot swap) are implemented to achieve resilience requirements in normal and adverse situations.	PR.PT-5: Mechanisms (e.g., failsafe, load balancing, hot swap) are implemented to achieve resilience requirements in normal and adverse situations	•	•	•	•	•	•	•	•	•	•	••	•	

Framework Subcategories		Access	Diagnostics	Disparities	Engagement	Ethics	Improvement	Legal	Outcomes	Optimize	Research	Secure	Transparency	Mapped to One or More Consider-ations?
Privacy Framework	Cybersecurity Framework													
	DE.AE-1: A baseline of network operations and expected data flows for users and systems is established and managed	●●●	●●●	●●	●●	●●	●●●	●●	●●	●●●	●●●	●●●	●●	
	DE.AE-2: Detected events are analyzed to understand attack targets and methods	●●	●●	●	●	●	●●	●	●	●●	●●	●●	●	
	DE.AE-3: Event data are collected and correlated from multiple sources and sensors	●●	●●	●	●	●	●●	●	●●	●●	●●	●●	●	
	DE.AE-4: Impact of events is determined	●●●	●●●	●●	●●	●●	●●●	●●	●●●	●●●	●●●	●●●	●●	YES
	DE.AE-5: Incident alert thresholds are established	●●	●●	●●	●●	●●	●●●	●●	●●	●●	●●●	●●●	●●	
	DE.CM-1: The network is monitored to detect potential cybersecurity events	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	

Framework Subcategories		Access	Diagnostics	Disparities	Engagement	Ethics	Improvement	Legal	Outcomes	Optimize	Research	Secure	Transparency	Mapped to One or More Consider-ations?
Privacy Framework	Cybersecurity Framework													
	DE.CM-2: The physical environment is monitored to detect potential cybersecurity events	●●	●●	●●	●●	●●	●●	●●	●●	●●	●●	●●	●●	
	DE.CM-3: Personnel activity is monitored to detect potential cybersecurity events	●●●	●●●	●●	●●	●●	●●●	●●	●●	●●●	●●●	●●●	●●	
	DE.CM-4: Malicious code is detected	●●●	●●●	●●	●●	●●	●●●	●●	●●	●●●	●●●	●●●	●●	
	DE.CM-5: Unauthorized mobile code is detected	●●	●●	●	●	●	●●	●	●●	●●	●●	●●	●	
	DE.CM-6: External service provider activity is monitored to detect potential cybersecurity events	●●●	●●●	●●	●●	●●	●●●	●●	●●	●●●	●●●	●●●	●●	
	DE.CM-7: Monitoring for unauthorized personnel, connections, devices,	●●●	●●●	●●	●●	●●	●●●	●●	●●	●●●	●●●	●●●	●●	

Framework Subcategories		Access	Diagnostics	Disparities	Engagement	Ethics	Improvement	Legal	Outcomes	Optimize	Research	Secure	Transparency	Mapped to One or More Consider-ations?
Privacy Framework	Cybersecurity Framework													
	and software is performed													
	DE.CM-8: Vulnerability scans are performed	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	YES
	DE.DP-1: Roles and responsibilities for detection are well defined to ensure accountability	●	●	●	●	●	●	●	●	●	●	●●	●	
	DE.DP-2: Detection activities comply with all applicable requirements	●●	●●	●	●●	●	●	●●●	●●	●	●●	●●●	●●	
	DE.DP-3: Detection processes are tested	●●	●●	●●	●●	●●	●	●●	●●	●●	●	●●	●●	
	DE.DP-4: Event detection information is communicated	●	●	●●	●●	●●	●	●●	●●	●	●	●	●●	
	DE.DP-5: Detection processes are continuously improved	●	●	●	●	●	●	●	●	●	●	●●	●	

Framework Subcategories		Access	Diagnostics	Disparities	Engagement	Ethics	Improvement	Legal	Outcomes	Optimize	Research	Secure	Transparency	Mapped to One or More Consider-ations?
Privacy Framework	Cybersecurity Framework													
	RS.RP-1: Response plan is executed during or after an incident	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	●●●	YES
	RS.CO-1: Personnel know their roles and order of operations when a response is needed	●●	●●	●●●	●●●	●●●	●●	●●●	●●●	●●	●●	●●	●●●	
	RS.CO-2: Incidents are reported consistent with established criteria	●●	●●●	●●●	●●●	●●●	●●	●●●	●●●	●●	●●	●●	●●●	YES
	RS.CO-3: Information is shared consistent with response plans	●●	●●	●●	●●	●●	●●	●●	●●	●●	●●	●●	●●	
	RS.CO-4: Coordination with stakeholders occurs consistent with response plans	●●	●●	●●	●●	●●	●●	●●	●●	●●	●●	●●	●●	
	RS.CO-5: Voluntary information sharing occurs with external stakeholders to achieve	●	●	●	●	●	●	●	●	●	●	●	●	

Framework Subcategories		Access	Diagnostics	Disparities	Engagement	Ethics	Improvement	Legal	Outcomes	Optimize	Research	Secure	Transparency	Mapped to One or More Consider-ations?
Privacy Framework	Cybersecurity Framework													
	broader cybersecurity situational awareness													
	RS.AN-1: Notifications from detection systems are investigated	●●	●●	●●	●●	●●	●●	●●	●●	●●	●	●●	●●	
	RS.AN-2: The impact of the incident is understood	●●●	●●●	●●	●●	●●	●●	●●	●●	●●●	●●	●●●	●●	
	RS.AN-3: Forensics are performed	●●●	●●●	●●	●●	●●	●	●●	●●	●●●	●	●●●	●●	
	RS.AN-4: Incidents are categorized consistent with response plans	●●●	●●●	●●	●●	●●	●	●●	●●	●●●	●	●●●	●●	
	RS.AN-5: Processes are established to receive, analyze and respond to vulnerabilities disclosed to the organization from internal and external sources (e.g. internal testing,	●●	●●	●	●	●	●	●	●●	●	●	●●	●	

Framework Subcategories		Access	Diagnostics	Disparities	Engagement	Ethics	Improvement	Legal	Outcomes	Optimize	Research	Secure	Transparency	Mapped to One or More Consider-ations?
Privacy Framework	Cybersecurity Framework													
	security bulletins, or security researchers)													
	RS.MI-1: Incidents are contained	●●●	●●●	●●	●●	●●	●	●●	●●	●●●	●	●●●	●●	
	RS.MI-2: Incidents are mitigated	●●●	●●●	●●	●●	●●	●●	●●	●●●	●●●	●	●●●	●●	
	RS.MI-3: Newly identified vulnerabilities are mitigated or documented as accepted risks	●●	●●	●●	●●	●●	●	●●	●●	●●	●	●●	●●	
	RS.IM-1: Response plans incorporate lessons learned	●●	●●	●●	●●	●●	●●	●●	●●	●●	●●	●●	●●	
	RS.IM-2: Response strategies are updated	●●	●●	●●	●●	●●	●●	●●	●●	●●	●●	●●	●●	
	RC.RP-1: Recovery plan is executed during or	●●●	●●●	●●	●●	●●	●●	●●	●●●	●●●	●●	●●●	●●	

Framework Subcategories		Access	Diagnostics	Disparities	Engagement	Ethics	Improvement	Legal	Outcomes	Optimize	Research	Secure	Transparency	Mapped to One or More Consider-ations?
Privacy Framework	Cybersecurity Framework													
	after a cybersecurity incident													
	RC.IM-1: Recovery plans incorporate lessons learned	●●	●●	●	●	●	●●	●	●●	●●	●●	●●	●	
	RC.IM-2: Recovery strategies are updated	●	●	●	●	●	●●	●	●●	●	●●	●●	●	
	RC.CO-1: Public relations are managed	●●	●●	●●	●●	●	●●	●	●●	●	●●	●●	●●	
	RC.CO-2: Reputation is repaired after an incident	●	●	●	●	●	●	●	●	●	●	●	●	
	RC.CO-3: Recovery activities are communicated to internal and external stakeholders as well as executive and management teams	●●	●●	●●	●●	●●	●●	●●	●●	●●	●●	●●●	●●	