S2 Appendix. Health Equity Across the AI Lifecycle (HEAAL).

Health Equity Across the AI Lifecycle (HEAAL)

A Framework for Healthcare Delivery Organizations to Mitigate the Risk of AI Solutions Worsening Health Inequities

health ai partnership

healthaipartnership.com

Table of Contents

Acknowledgement

We thank the Gordon and Betty Moore Foundation for funding this project, Health AI Partnership leadership council and coordinating center for leading this project, framework developers for sharing their expertise, NewYork-Presbyterian (NYP) and Parkland Center for Clinical Innovation (PCCI) for presenting the case studies for the workshop, experts discussants for sharing reflections on the case studies, and workshop participants for sharing their insights, and Duke Heart Center administrators for grant management.

Introduction

While AI adoption in health systems is increasing, there is no formal guidance on how healthcare delivery organizations can ensure that AI does not worsen health inequities. This framework is designed to systematically promote health equity when a healthcare delivery organization is considering adopting a new solution that uses AI. It defines specific procedures that healthcare delivery leaders and project leaders can use to examine the potential impact of a new AI solution on health equity and make evidence-based decisions throughout the AI lifecycle.

Health AI Partnership is

A multi-stakeholder collaborative that seeks to empower healthcare organizations to use AI safely, effectively, and equitably.

The trusted resource for contemporary guidance for healthcare professionals using AI and related emerging technologies.

The platform for community-generated, expert-curated guidance, resources, and standards for responsible use of AI in healthcare.

A network that creates safe spaces for peer learning and collaboration to address the most challenging issues health care leaders face while implementing AI.

Mission

Empower healthcare professionals to use AI effectively, safely, and equitably through community-informed up-to-date standards.

Vision

Be the trusted partner and up-to-date source of actionable guidance for healthcare proffesionals using AI.

Values

Advance health equity
Prioritize solutions that advance health equity and eliminate the AI digital device.

Improve patient care
Ensure that AI adoption is driven by patient care needs, not technical novelty.

Improve the workplace
Surface social-technical challenges in AI use and foster a positive work environment.

Build community
Create safe spaces to share learnings and consult peers.

How was the HEAAL developed?

The framework was developed through a case study workshop involving leaders from healthcare delivery organizations and ecosystem partner sites. Three innovation teams were recruited to present case studies. Seventy-seven representatives from ten healthcare organizations and four ecosystem partners shared their AI adoption experiences.

Six framework developers from diverse backgrounds —a clinician, a community representative, a computer scientist, a legal and regulatory expert, a project manager, and a sociotechnical scholar—created and refined the framework structure. Eight Health AI Partnership leaders evaluated the framework and provided feedback. Design researchers facilitated the design process, synthesizing key insights, conducting two rounds of usability testing, and refining the framework.

How to use the HEAAL?

Step 1

Review the equity principles outlined on page 9.

Step 2

Select an adoption stage along with its corresponding key decision point from the overview provided on page 12.

Step 3

Scroll through procedures in each Key Decision Point to explore the full details.

Step 4

To explore a different Key Decision Point, return to the overview by clicking on the OVERVIEW button in the top right corner.

Navigation

OVERVIEW TABLE OF CONTENTS

At any time, use the navigation buttons located in the top right corner of the HEAAL to return to the preceding sections.

What's inside?

Adoption Stages

The framework extends across four stages of AI lifecycle. Each stage is distinguished by its own color bar:

- Problem identification and procurement
- Development and adaptation
- Clinical integration
- Lifecycle management

Key Decisions Points

The HEAAL is structured around the Health AI Partnership's Eight Key Decision Points of AI adoption.

Key Decisions Points

1. Identify and prioritize a problem
2. Define AI product scope and intended us
3. Develop success measures
4. Define AI product scope and intended use
5. Generate evidence of safety, efficacy and equity
6. Execute AI solution roll out
7. Monitor the AI solution
8. Update or decommission the AI solution

Procedures

The framework includes procedures that should be tested across the Key Decision Points to promote and advance health equity. All procedures are described in the HEAAL Chapter.

The procedures complement and augment the Key Decision Point topic guides that are published on the Health AI Partnership website.

Procedures for Existing vs. New Solutions

While all procedures should be considered for all AI solutions of interest, some procedures are tested in different decision points or in a different sequential order, depending on whether the solutions already exist or not.

- For evaluating an existing AI solution, follow procedures written in: red and **black text**.
- For evaluating a new AI solution, follow procedures written in: blue and **black text**.

Footnotes

Throughout the framework, footnotes are embedded to provide additional information and examples for some specific procedures.

Equity principles that the HEAAL assesses

Accountability

Ensure that potential adverse impacts of using the solution are overseen by specific stakeholders who have clear responsibilities.

Fairness

Establish and evaluate meaningful fairness criteria that can empower the healthcare delivery organization to track progress and identify problems. Ensure that the solution performs equitably across disadvantaged patient subgroups.

Reliability and validity

Ensure that the solution performance is reliable and valid.

Fitness for purpose

Ensure that the proposed solution solves the identified problem. There should be a well-specified intended use statement, a comprehensive understanding of user needs, and an assessment differentiating the solution from alternative approaches to solve the problem.

Transparency

Communicate the processes of model development, implementation, potential risks, and harms associated with the model use.

Stakeholders involved in completing the procedures of the HEAAL

S

Strategic
Stakeholders who develop strategic plans and make decisions that align with organizational interests

O

Operational
Stakeholders who manage workflow and make decisions to integrate

C

Clinical
Stakeholders who provide clinical care to patients

T

Technical
Stakeholders who develop the model and its infrastructure

R

Regulatory
Stakeholders who review the model from regulatory, compliance and ethical perspectives

P

Patient
Stakeholders who receive clinical care and provide insights on their community experiences

CC

Clinical Champion
Clinical stakeholders who lead the project and provide clinical expertise in model development

Sources of data used to complete the procedures of the HEAAL

Local healthcare retrospective data

Historical healthcare data that is curated within the primary healthcare delivery organization seeking to adopt an AI product. The local data can be sourced from a variety of systems, including the EHR, radiology PACS system, medical claims, audit logs, electrocardiograms, and high-frequency vital sign monitors. When a model is internally developed, the local healthcare retrospective data set is used for training the model

Local healthcare prospective data

Real-time healthcare data that is curated within the primary healthcare delivery organization seeking to adopt an AI product. The local data can be sourced from a variety of systems, including the EHR, radiology PACS system, medical claims, audit logs, electrocardiograms, and high-frequency vital sign monitors. The local healthcare prospective data set is used for validating a model during a 'silent trial' and for using the model in clinical care.

Training data

Data used for training a model. When the model is externally developed, the training data set contains data from an external source. When the model is internally developed, the training data set is sourced from local healthcare retrospective data.

Local non-healthcare data

Non-healthcare data that is curated within a geographic setting where a healthcare delivery organization is based. The local non-healthcare data can be derived from a variety of external sources, including US Census.

Literature review

Data collected through reviewing previously published scholarly works on a specific topic.

Organizational data

Data that describes central characteristics of organizations, their internal structures, processes, and behavior as corporate actors in different social and economic contexts. The organizational data includes Key performance Indicators (KPIs) that quantify progress toward an intended result. KPIs provide a focus for strategic and operational improvement and create an analytical basis for decision making.

Qualitative data

Data collected through qualitative research methods, including surveys, focus groups, and interviews.

Overview of the HEAAL

Problem identification and procurement

1. Identify and prioritize a problem

S O C P

Flag for an inequitably prioritized problem. Then, identify disadvantaged patient subgroups.

2. Define AI product scope and intended use

S O C T R P

Flag for representation bias.

+

FOR EXISTING SOLUTION:
Flag for representation bias, label choice bias, and hidden stratification. Then compare model performance across all patient subgroups and flag concerns.

Development and adaptation

3. Develop success measures

S O C

Establish equity objectives and fairness metrics.

4. Define AI product scope and intended use

S O C R P

Flag for insufficient engagement from end users and members of disadvantaged patient subgroups, a lack of inclusivity in the solution design and the potential bias in end users.

5. Generate evidence of safety, efficacy and equity

O C T P

Flag for missing data. Compare model performance across all patient subgroups and flag model performance not aligning with equity objectives and fairness metrics.

+

FOR NEW SOLUTION:
Flag for Representation bias, label choice bias, and hidden stratification.

Clinical integration

6. Execute AI solution roll out

S O C R P

Operationalize the communication plan and gather feedback.

Lifecylce management

7. Monitor the AI solution

O T P

Flag for health inequities and monitoring outcomes not aligning with equity objectives and fairness metrics.

8. Update or decommission the AI solution

O C T P

Make decisions regarding updating decommissioning, or expanding the solution.

STAKEHOLDERS: S Strategic · O Operational · C Clinical · T Technical · R Regulatory · P Patient · CC Clinical Champion

EQUITY PRINCIPLES: Accountability · Fitness for purpose · Reliability and validity · Fairness · Transparency

SOURCES OF DATA: Local healthcare retrospective data · Local healthcare prospective data · Training data · Literature review · Local non-healthcare data · Qualitative data · Organizational data

Health Equity Across the AI Lifecycle (HEAAL)

1. Identify and prioritize a problem

Decision Point 1 focuses on identifying and prioritizing problems within health care delivery organizations and surfacing stakeholders affected by problems. It also describes how to determine different dimensions of problems and assess the suitability of AI as a technical approach to address problems.

a **Ensure that problems are prioritized and funded equitably across all patient subgroups.**

S

i. Prioritize problems that demonstrate potential to positively impact all patient subgroups.

ii. Ensure to account for the number and types of affected patients in making funding and resource allocation decisions.

b **Determine whether there are patient populations for whom a solution to the prioritized problem should not be used, should be used differently, or whose experience with the system should be closely monitored.**

C O P

i. Review literature on epidemiology and health disparities to understand the current state of health inequities within the context of the prioritized problem.[1]

ii. Identify social determinants of health (SDOH) and demographic subgroups who may experience health inequities (i.e., disadvantaged patient subgroups) within the context of the prioritized problem.[2]

iii. Continue identifying health inequities and disadvantaged patient subgroups by interviewing or surveying personnel and patient community members best positioned to understand the experiences of disadvantaged patient subgroups.[3]

iv. Drawing from information collected from the previous procedures, establish a list of health inequities and disadvantaged patient subgroups. Make sure to list intersecting identities that may be associated with worse inequities and include references.

1. Understanding the current state is important for ensuring that the proposed solution accounts for contextual characteristics of patients that may shape the inequities so that all patients can receive equitable care. Review Table 2 for examples of different types of health inequities.
2. It is important to note that the absence of documented inequalities does not mean the absence of inequalities. For example, some demographic subgroups, such as transgender and gender-nonconforming individuals and undocumented immigrants, have poor data collection to even understand inequities. (Chen IY, Pierson E, Rose S, Joshi S, Ferryman K, Ghassemi M. Ethical Machine Learning in Healthcare. Annu Rev Biomed Data Sci. 2021;4(1):1-22. doi:10.1146/annurev-biodatasci-092820-114757) While information about language that patients speak may not have been captured, language serves as an important social identity of patients.
3 .Such personnel may include healthcare providers, social workers, patient navigators, and patient community members identified as negatively impacted by health inequities. Sample questions to ask the personnel are described in the worksheet.

STAKEHOLDERS: S Strategic; O Operational; C Clinical; T Technical; R Regulatory; P Patient; CC Clinical Champion

EQUITY PRINCIPLES: Accountability; Fitness for purpose; Reliability and validity; Fairness; Transparency

SOURCES OF DATA: Local healthcare retrospective data; Local healthcare prospective data; Training data; Literature review; Local non-healthcare data; Qualitative data; Organizational data

2. Define AI product scope and intended use

Decision Point 2 describes how organizations can assess the feasibility and viability of adopting AI to solve problems. It also explains how to conduct preliminary assessments of AI products, assess legal risks of AI product adoption, and audit the process by which AI product investments are made.

a **List alternative solutions for the problem, including non-technical interventions and other non-AI technical interventions.**

C O T

b **Define an ideal label for model development.**

C O P T

i. Conduct interviews, focus groups, or surveys with a cross-functional team to examine how patients should be identified by an algorithm to optimize decision making within the context of the prioritized problem. Allow the cross-functional team to assume that the algorithm had access to pristine and complete information about all patients.[4]

ii. Assess responses from the cross-functional team and identify the most promising clinical outcome used for model development which will be defined as an ideal label. Ensure to document the rationale for selecting the label.

4. Individuals and groups who understand the experiences of disadvantaged patient subgroups can provide important context around the labels and suggest a desirable state that could be pursued using the algorithm. (Mccradden M, Odusi O, Joshi S, et al. What's fair is... fair? Presenting JustEFAB, an ethical framework for operationalizing medical ethics and social justice in the integration of clinical machine learning. 2023 ACM Conf Fairness, Account, Transpar. Published online 2023:1505-1519. doi:10.1145/3593013.3594096)

STAKEHOLDERS: S Strategic; O Operational; C Clinical; T Technical; R Regulatory; P Patient; CC Clinical Champion

EQUITY PRINCIPLES: Accountability; Fitness for purpose; Reliability and validity; Fairness; Transparency

SOURCES OF DATA: Local healthcare retrospective data; Local healthcare prospective data; Training data; Literature review; Local non-healthcare data; Qualitative data; Organizational data

c Seek an approval from an institutional review board, ethical review board, or research ethics board to access and use local healthcare retrospective data.

R

d Assess health inequities present in the local healthcare retrospective data and identify disadvantaged patient subgroups within the context of the prioritized problem.

T

i. Specify which data elements can be used to measure health inequities identified in 1(b)(iv).

ii. Using the identified data elements, examine whether health inequities identified in 1(b)(iv) are present within local healthcare retrospective data.

iii. Combine information collected from the previous procedure 2(d)(ii) with information collected from 1(b)(iv) to establish a list of health inequities and disadvantaged patient subgroups within the local healthcare delivery setting. [5] The list should include references and supporting data. All other patient subgroups, not including the disadvantaged patient subgroups, are defined as advantaged patient subgroups.[6]

e Examine whether a local healthcare retrospective data set is representative of demographic representation of local non-healthcare data.

T

i. Conduct a demographic analysis of the patient cohort within the local healthcare retrospective data set.

ii. Examine whether the demographics of the local healthcare retrospective data reflects demographic representation of local non-healthcare data derived from external data sources.[7]

iii. If demographics of the local healthcare retrospective data does not significantly align with demographics of the local non-healthcare data, flag for potential representation bias.

5. For example, if a patient subgroup A was identified as a disadvantaged patient subgroup in 1(b)(iv) and a patient subgroup B was identified as a disadvantaged patient subgroup in 2(d)(ii), both patient subgroups A and B should be considered as disadvantaged patient subgroups.
6. Keep in mind that the list of disadvantaged patient subgroups and advantaged patient subgroups may be updated as the analysis progresses.
7. For example, US Census demographic information can be used to curate demographic information about the population within a catchment area of a healthcare delivery organization.

STAKEHOLDERS: S Strategic; O Operational; C Clinical; T Technical; R Regulatory; P Patient; CC Clinical Champion

EQUITY PRINCIPLES: Accountability; Fitness for purpose; Reliability and validity; Fairness; Transparency

SOURCES OF DATA: Local healthcare retrospective data; Local healthcare prospective data; Training data; Literature review; Local non-healthcare data; Qualitative data; Organizational data

FOR EXISTING SOLUTION

If considering an existing AI solution, proceed to the next set of procedures written in red to assess AI solution options and select the most optimal one.
If an AI solution does not exist and is being considered for internal development, skip to Decision Point 3.

f

Assess health inequities present in the model training data and identify disadvantaged patient subgroups within the context of the prioritized problem.

T

i. Using data elements that are similar to the ones identified in 2(d)(i), examine whether health inequities identified in 1(b)(iv) are present within the training data.

ii. Confirm the presence or absence of health inequities among disadvantaged patient subgroups previously identified in 2(d)(iii).

iii. The implication of the presence or absence of health inequities within the training data depends on whether health inequities were identified within local healthcare retrospective data in 2(d)(ii):
- If health inequities were present within local healthcare retrospective data, the presence or absence of health inequities within the training data conveys little to no concern for worsening existing health inequities in the local setting.
- If health inequities were not present within local healthcare retrospective data, the absence of health inequities within the training data conveys little concern for creating future health inequities. However, the presence of health inequities within the training data conveys potential concern for creating future health inequities.

g

Examine whether the model training data is representative of the demographics present within the local healthcare retrospective data.

T

i. For each AI solution, conduct a demographic analysis of the patient cohort within the training data.

ii. Examine whether the training data reflects a demographic representation of the local healthcare retrospective data analyzed in 2(e)(i). [8]

iii. If demographics of the training data does not significantly align with demographics of the local healthcare retrospective data, flag for potential representation bias. Prioritize AI solutions that are trained on data sets that are representative of the patient subgroups within the local healthcare context.

h

Analyze label choice bias across disadvantaged and advantaged patient subgroups.

C O P T

i. For each AI solution, gather a list of the actual labels used for model training.

ii. From the list, select a closer-to-ideal label, even if it is available only for a small subset of patients. Use this label for analysis in the next procedures.

iii. Consult with a cross-functional team and assess how well the actual label aligns with the ideal label identified in 2(b)(ii) across disadvantaged and advantaged patient subgroups identified in 2(d)(iii).

iv. If the actual label does not accurately capture the ideal label, flag for label choice bias. Prioritize AI solutions with actual labels that accurately measure the ideal label.

8. For example, if there is a minimal representation of patient subgroups identified in section 1(b)(iv) in the training data, this representation should be a big red flag.

STAKEHOLDERS: S Strategic; O Operational; C Clinical; T Technical; R Regulatory; P Patient; CC Clinical Champion

EQUITY PRINCIPLES: Accountability; Fitness for purpose; Reliability and validity; Fairness; Transparency

SOURCES OF DATA: Local healthcare retrospective data; Local healthcare prospective data; Training data; Literature review; Local non-healthcare data; Qualitative data; Organizational data

i **Ensure that the model features are relevant to its actual label and capture the same meanings across disadvantaged and advantaged patient subgroups.**

C T

i. For each AI solution, gather the actual labels and a comprehensive list of model features used for model training.

ii. Ensure that all model features accurately represent concepts relevant to the actual label.

iii. Consult with clinicians and understand if there are known differences in how model features are gathered and represented across disadvantaged and advantaged patient subgroups identified in 2(d)(iii).[9]

iv. For each feature, examine how it is correlated with actual labels across all patient subgroups and identify whether this association manifests inconsistently in disadvantaged patient subgroups identified in 2(d)(iii).

v. If the association between model features and the actual label is inconsistent across different patient subgroups, flag for measurement bias. Prioritize AI solutions that use features that have consistent meanings and are captured robustly across all patient subgroups.

j **Identify potential hidden stratification that masks unequal model performance between disadvantaged and advantaged patient subgroups.**

T

i. Identify diagnostic and treatment subgroups within the actual label.[10]

ii. Compare model performance between the identified subgroups.

iii. If the model performance of a subgroup with diagnosis and treatment is better than the model performance of a subgroup without diagnosis and treatment, flag for *hidden stratification*.[11] Consider alternative solutions (return to 2(a)) or retrain the model after excluding a cohort of patients from the subgroup with diagnosis and treatment.

iv. If the model performance of a subgroup without diagnosis and treatment is better than the model performance of a subgroup with diagnosis and treatment, or if model performance between the subgroups is similar, then compare model performance between disadvantaged and advantaged patient subgroups defined in 2(d)(iii) within each diagnostic and treatment subgroup.

v. Prioritize AI solutions that perform equally well on disadvantaged and advantaged patient subgroups within each diagnostic and treatment subgroup.

k **Gather model performance data and compare it between disadvantaged and advantaged patient subgroups.**

T

i. For each AI solution, gather model performance data for disadvantaged patient subgroups and advantaged patient subgroups identified in 2(d)(iii).

ii. Compare the model performance of disadvantaged patient subgroups to the model performance of advantaged patient subgroups.

iii. If model performance is worse for disadvantaged patient subgroups than model performance for advantaged patient subgroups, the implication depends on whether health inequities were identified in 2(d)(ii):
- If health inequities were confirmed, the difference in model performance could be due to existing health inequities, not due to modeling, and thus, conveys minimal concern with the model.
- If health inequities were not confirmed, the difference in model performance could be due to modeling and create future health inequities and thus, conveys potential concern with the model.

9. For example, examine whether there are any known symptoms that manifest differently for different patient demographics.
10. For example, diagnostic and treatment subgroups may include (1) a subgroup that has completed all relevant prior workup versus a subgroup who has not (workup inequity), (2) a subgroup that is diagnosed at an earlier disease state versus a subgroup that is diagnosed at a later disease state (diagnosis inequity), (3) a subgroup that receives important interventions prior to or immediately after diagnosis versus a subgroup that does not receive important interventions (treatment inequity), and (4) a subgroup with poor outcomes versus a subgroup with good outcomes (outcome inequity).
11. Oakden-Rayner L, Dunnmon J, Carneiro G, Ré C. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In: Vol 52. ACM CHIL '20: ACM Conference on Health, Inference, and Learning. ; 2020:151-159. doi:10.1145/3368555.3384468

STAKEHOLDERS: S Strategic; O Operational; C Clinical; T Technical; R Regulatory; P Patient; CC Clinical Champion

EQUITY PRINCIPLES: Accountability; Fitness for purpose; Reliability and validity; Fairness; Transparency

SOURCES OF DATA: Local healthcare retrospective data; Local healthcare prospective data; Training data; Literature review; Local non-healthcare data; Qualitative data; Organizational data

k

iv. If model performance for disadvantaged patient subgroups is similar to model performance for advantaged subgroups, the implication depends on whether health inequities were identified in 2(d)(ii):
- If health inequities were confirmed, the similarity in model performance could be due to modeling and further reinforce existing health inequities and thus, conveys potential concern with the model.
- If health inequities were not confirmed, the similarity in model performance could be due to the absence of health inequities and thus, conveys minimal concern with the model.

v. Prioritize AI solutions with minimal concern.[12]

l

Determine which SDOH and demographic data are appropriate to be included in the model to minimize potential risk of worsening health inequities.

T

i. For each AI solution, review the comprehensive list of model features gathered in 2(h)(i).

ii. Identify solutions that use patient SDOH and demographic data as model features.

iii. For solutions that use patient SDOH and demographic data as model features, gather a rationale for inclusion of each SDOH and demographic feature.

iv. Gather model performance data without using SDOH and demographic data as model features.

v. Compare model performance data gathered in the previous procedure 2(l)(iv) with model performance data with SDOH and demographic features gathered in 2(k)(i) across disadvantaged and advantaged patient subgroups identified in 2(d)(iii).[13]

vi. Prioritize AI solutions that use SDOH and demographic data to improve performance for disadvantaged patient subgroups identified in 2(d)(iii). However, if model performance remains robust across disadvantaged patient subgroups without using SDOH and demographic data, minimize use of this data.

m

Determine which potential solution best solves the problem for disadvantaged patient subgroups.

C O T S

i. Select a solution, which may or may not use AI, that best addresses the prioritized problem for disadvantaged patient subgroups identified in 2(d)(iii).

ii. If an AI solution is selected, proceed to Decision Point 3 and onward.

12. For patient subgroups identified in section 2(d)(iii) who are poorly represented in the local context (e.g., Native American women in Boston), AI solution performance may have wide confidence bounds. Ensure that performance measures account for uncertainty.
13. Sometimes, removing SDOH and demographic data as model features may decrease model performance. For example, removing race worsened model performance (Khor S, Haupt EC, Hahn EE, Lyons LJL, Shankaran V, Bansal A. Racial and Ethnic Bias in Risk Prediction Models for Colorectal Cancer Recurrence When Race and Ethnicity Are Omitted as Predictors. JAMA Netw Open. 2023;6(6):e2318495. doi:10.1001/jamanetworkopen.2023.18495).

STAKEHOLDERS: S Strategic; O Operational; C Clinical; T Technical; R Regulatory; P Patient; CC Clinical Champion

EQUITY PRINCIPLES: Accountability; Fitness for purpose; Reliability and validity; Fairness; Transparency

SOURCES OF DATA: Local healthcare retrospective data; Local healthcare prospective data; Training data; Literature review; Local non-healthcare data; Qualitative data; Organizational data

3. Develop success measures

Decision Point 3 focuses on defining the scope of use, constraints, and dependencies of AI products. It also describes how to define technical performance targets for AI products as well as measures of success for AI products used in practice.

a **Establish equity objectives for implementation of the selected AI solution.**

C

i. Review baseline inequities calculated in 2(d)(ii).

ii. For each disadvantaged and advantaged patient subgroup identified in 2(d)(iii), set objectives for AI solution implementation to address health inequities identified in 2(d)(ii) in terms of health and economic outcomes.[14] Make sure to establish a cut-off or threshold for considering when to pause use or decommission the solution.

b **Identify the most appropriate fairness metrics to use for the selected AI product and its design**

C O T

i. Review the context of the use case and disadvantaged patient subgroups identified in 2(d)(iii).

ii. Review literature and identify fairness metrics relevant to the solution.

iii. Discuss which fairness metrics can be used to achieve equity objectives defined in 3(a)(ii).

iv. Discuss which fairness metrics are pragmatic and align with health system priorities.

v. Establish the most appropriate fairness metrics for the use case and document the rationale for selecting the particular fairness metrics.[15] Make sure to establish a cut-off or threshold for considering when to decommission the solution.

14. The equity objectives may range from maintaining the current level of inequity to significantly reducing it.
15. Choices about fairness metrics should be revisable based on the clinical evidence observed through real-time model use and an ongoing monitoring process (Mccradden M, Odusi O, Joshi S, et al. What's fair is...fair? Presenting JustEFAB, an ethical framework for operationalizing medical ethics and social justice in the integration of clinical

STAKEHOLDERS: S Strategic; O Operational; C Clinical; T Technical; R Regulatory; P Patient; CC Clinical Champion

EQUITY PRINCIPLES: Accountability; Fitness for purpose; Reliability and validity; Fairness; Transparency

SOURCES OF DATA: Local healthcare retrospective data; Local healthcare prospective data; Training data; Literature review; Local non-healthcare data; Qualitative data; Organizational data

4. Define AI product scope and intended use

Decision Point 4 describes how organizations design and test optimal workflows for clinician-facing AI products. It also explains how to adapt pre-existing operational structures, workflows, and technologies to enable successful integration of AI products.

a **Ensure that the solution design is informed by meaningful and pragmatic recommendations from members of disadvantaged patient subgroups.**

P

i. Recruit patients and patient advocates with lived experience as members of disadvantaged patient subgroups identified in 2(d)(iii).

ii. Conduct interviews or focus groups with patients and patient advocates to surface problems and concerns they have for the solution design. Ensure to seek input from them about the various forms of support disadvantaged patient subgroups may need to benefit the most from using the solution.

b **Ensure that the solution design promotes inclusivity of clinical end-users and usability of the solution**

C

i. Recruit a representative sample of clinical end-users to test accessibility, inclusivity, and usability of the solution.

ii. Conduct interviews, focus groups, or surveys with clinical end-users to surface user needs and concerns they have for the solution design. Ensure to seek input from them about the various forms of support they may need to use the solution efficiently.

iii. Ensure to design the solution in a way that minimizes barriers to effective use for diverse clinical end-users.[16]

16. For example, issues like color blindness, sensitivity to light levels, and other ergonomic factors that may prevent end-users with disabilities from using the solution should be addressed (Mccradden M, Odusi O, Joshi S, et al. What's fair is... fair? Presenting JustEFAB, an ethical framework for operationalizing medical ethics and social justice in the integration of clinical machine learning. 2023 ACM Conf Fairness, Account, Transpar. Published online 2023:1505-1519. doi:10.1145/3593013.3594096).

STAKEHOLDERS: S Strategic; O Operational; C Clinical; T Technical; R Regulatory; P Patient; CC Clinical Champion

EQUITY PRINCIPLES: Accountability; Fitness for purpose; Reliability and validity; Fairness; Transparency

SOURCES OF DATA: Local healthcare retrospective data; Local healthcare prospective data; Training data; Literature review; Local non-healthcare data; Qualitative data; Organizational data

c **Design complementary non-technical solution components required to achieve equity objectives for implementation of the solution.**

C O S

i. Identify complementary non-technical solution components and resources that will be required to achieve each equity objective defined in 3(a)(ii).[17] Ensure to incorporate recommendations received in 4(a) and 4(b) into designing non-technical interventions.

ii. Evaluate internal resources and allocate necessary resources for implementing non-technical solution components.

iii. Develop training material and communication plan to equip frontline clinicians to achieve equity objectives defined in 3(a)(ii).

d **Align clinical end-users and organizational leaders to achieve equity objectives.**

C O R S

i. Communicate the equity objectives defined in 3(a)(ii) and their importance to organizational leaders. Ensure that the equity objectives are not presented as secondary to clinical or operational objectives.

ii. Seek support from organizational leaders to implement necessary non-technical solution components defined in 4(c)(i).

iii. Work with management and legal to establish incentives for frontline clinicians involved in the AI solution implementation.[18]

17. Anticipate that AI solutions need to be complemented with other resources, workflows, and capabilities to effectively address inequities. Examples of non-technical solution components for the various categories of inequity may include:
Ensuring sufficient capabilities, such as personnel and equipment, to perform diagnostic interventions or procedures for affected patient subgroups (workup inequity)
Improving access to diagnostic testing and evaluation for affected patient subgroups (diagnosis inequity)
Ensuring sufficient capabilities and access to treatments and interventions for affected patient subgroups (treatment inequity
Enhancing monitoring and sufficient capabilities to effectively manage affected patient subgroups (outcome inequity)
18. Incentives may include public recognition for meeting performance targets, inclusion in promotion criteria, and possibly monetary awards under certain circumstances. Make sure that monetary awards related to effective use of AI do not induce or generate additional health service business, which may violate anti-kickback statutes.

STAKEHOLDERS: S Strategic; O Operational; C Clinical; T Technical; R Regulatory; P Patient; CC Clinical Champion

EQUITY PRINCIPLES: Accountability; Fitness for purpose; Reliability and validity; Fairness; Transparency

SOURCES OF DATA: Local healthcare retrospective data; Local healthcare prospective data; Training data; Literature review; Local non-healthcare data; Qualitative data; Organizational data

5. Generate evidence of safety, efficacy, and equity

Decision Point 5 focuses on local validation of AI products before clinical use and identifying foreseeable and unresolved risks resulting from AI use in clinical care. It also describes how to determine if AI products should be clinically integrated or abandoned.

a **Assess completeness and quality of local data required to construct model features across disadvantaged patient subgroups.**

C P T

i. Examine local healthcare retrospective data and assess missingness of all model features across disadvantaged and advantaged patient subgroups identified in 2(d)(iii).

ii. Identify the model features that are missing at different rates for disadvantaged patient subgroups identified in 2(d)(iii).

iii. For each model feature identified in 5(a)(ii), surface potential causes of differential missingness by consulting with a cross-functional team involved in data capture.

iv. If the model features need to be sourced differently for disadvantaged patient subgroups, make necessary changes.

v. Impute data in a fashion that minimizes inequities. If existing values within a disadvantaged patient subgroup identified in 2(d)(iii) differ substantially from values within an advantaged patient subgroup, consider imputation using subgroup statistics rather than full population statistics.

vi. If missingness of a model feature continues to differ substantially for disadvantaged patient subgroups identified in 2(d)(iii), consider including a missing indicator as a model feature.

b **Seek an approval from an institutional review board, ethical review board, or research ethics board to access and use local healthcare prospective data.**

R

i. Recruit a representative sample of clinical end-users to test accessibility, inclusivity, and usability of the solution.

ii. Conduct interviews, focus groups, or surveys with clinical end-users to surface user needs and concerns they have for the solution design. Ensure to seek input from them about the various forms of support they may need to use the solution efficiently.

iii. Ensure to design the solution in a way that minimizes barriers to effective use for diverse clinical end-users.

STAKEHOLDERS: S Strategic; O Operational; C Clinical; T Technical; R Regulatory; P Patient; CC Clinical Champion

EQUITY PRINCIPLES: Accountability; Fitness for purpose; Reliability and validity; Fairness; Transparency

SOURCES OF DATA: Local healthcare retrospective data; Local healthcare prospective data; Training data; Literature review; Local non-healthcare data; Qualitative data; Organizational data

FOR AN AI SOLUTION THAT IS BEING NEWLY DEVELOPED

For an AI solution that is being newly developed, proceed to the next set of procedures written in blue. For an AI solution that already exists, skip to 5(i).

c **Analyze label choice bias across disadvantaged and advantaged patient subgroups.**

C O P T

i. Examine a list of the actual labels used for model training.

ii. From the list, select a closer-to-ideal label, even if it is available only for a smallsubset of patients. Use this label for analysis in the next procedures.

iii. Consult with a cross-functional team and assess how well the actual label aligns with the ideal label identified in 2(d)(ii) across disadvantaged and advantaged patient subgroups identified in 2(d)(iii).

iv. Ensure that the AI solution uses actual labels that accurately measure the ideal label. If the actual label does not accurately capture the ideal label, the solution has a risk of label choice bias.

d **Ensure that the model features are relevant to its actual label and capture the same meanings across disadvantaged and advantaged patient subgroups.**

C T

i. Examine the actual labels and a comprehensive list of model features used for model training.

ii. Ensure that all model features accurately represent concepts relevant to the actual label.

iii. Consult with clinicians and understand if there are known differences in how model features are gathered and represented across disadvantaged patient subgroups identified in 2(d)(iii).[19]

iv. For each feature, examine how it is correlated with actual labels across all patient subgroups and identify whether this association manifests inconsistently in disadvantaged patient subgroups identified in 2(d)(iii).

v. Ensure that the AI solution uses features that have consistent meanings and are captured robustly across all patient subgroups. If the association between model features and the actual label is inconsistent across different patient subgroups, the solution has a risk of measurement bias.

e **Identify potential hidden stratification that masks unequal model performance between disadvantaged and advantaged patient subgroups.**

T

i. Identify diagnostic and treatment subgroups within the actual label.[20]

ii. Compare model performance between the identified subgroups.

iii. If the model performance of a subgroup with diagnosis and treatment is better than the model performance of a subgroup without diagnosis and treatment, flag for hidden stratification. [21] Consider alternative solutions (return to 2(a)) or retrain the model after excluding a cohort of patients from the subgroup with diagnosis and treatment.

iv. If the model performance of a subgroup without diagnosis and treatment is better than the model performance of a subgroup with diagnosis and treatment, or if model performance between the subgroups is similar, then compare model performance between disadvantaged and advantaged patient subgroups defined in 2(d)(iii) within each diagnostic and treatment subgroup.

v. Ensure that the AI solution performs equally well on disadvantaged and advantaged patient subgroups within each diagnostic and treatment subgroup.

19. For example, examine whether there are any known symptoms that manifest differently for differentpatient demographics.
20. For example, diagnostic and treatment subgroups may include (1) a subgroup that has completed all relevant prior workup versus a subgroup who has not (workup inequity), (2) a subgroup that is diagnosed at an earlier disease state versus a subgroup that is diagnosed at a later disease state (diagnosis inequity), (3) a subgroup that receives important interventions prior to or immediately after diagnosis versus a subgroup that does not receive important interventions (treatment inequity), and (4) a subgroup with poor outcomes versus a subgroup with good outcomes (outcome inequity).
21. Oakden-Rayner L, Dunnmon J, Carneiro G, Ré C. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In: Vol 52. ACM CHIL '20: ACM Conference on Health, Inference, and Learning. ; 2020:151-159. doi:10.1145/3368555.3384468

STAKEHOLDERS: S Strategic; O Operational; C Clinical; T Technical; R Regulatory; P Patient; CC Clinical Champion

EQUITY PRINCIPLES: Accountability; Fitness for purpose; Reliability and validity; Fairness; Transparency

SOURCES OF DATA: Local healthcare retrospective data; Local healthcare prospective data; Training data; Literature review; Local non-healthcare data; Qualitative data; Organizational data

f

Assess model performance and compare it between disadvantaged and advantagedpatient subgroups.

T

i. Assess model performance for disadvantaged patient subgroups and advantaged patient subgroups identified in 2(d)(iii).

ii. Compare the model performance of disadvantaged patient subgroups to the model performance of advantaged patient subgroups.

iii. If model performance is worse for disadvantaged patient subgroups than model performance for advantaged patient subgroups, the implication depends on whether health inequities were identified in 2(d)(ii):

- If health inequities were confirmed, the difference in model performance could be due to existing health inequities, not due to modeling, and thus, conveys minimal concern with the model.
- If health inequities were not confirmed, the difference in model performance could be due to modeling and create future health inequities and thus, conveys potential concern with the model.

iv. If model performance for disadvantaged patient subgroups is similar to model performance for advantaged subgroups, the implication depends on whether health inequities were identified in 2(d)(ii):

- If health inequities were confirmed, the similarity in model performance could be due to modeling and further reinforce existing health inequities and thus, conveys potential concern with the model.
- If health inequities were not confirmed, the similarity in model performance could be due to the absence of health inequities and thus, conveys minimal concern with the model.

v. If the model has potential concern, retrain the model to ensure that the model has minimal concern.[22]

g

Determine which SDOH and demographic data are appropriate to be included in the model to minimize potential risk of worsening health inequities.

T

i. Review the comprehensive list of model features gathered in 5(d)(i).

ii. If the solution uses patient SDOH and demographic data as model features, gather a rationale for inclusion of each SDOH and demographic feature.

iii. Assess model performance without using SDOH and demographic data as model features.

iv. Compare model performance data measured in the previous procedure 5(g)(iii) with model performance data measured in 5(f)(i) across disadvantaged and advantaged patient subgroups identified in 2(d)(iii).[23]

v. Use SDOH and demographic data to improve performance for disadvantaged patient subgroups identified in 2(d)(iii). However, if model performance remains robust across disadvantaged patient subgroups without using SDOH and demographic data, minimize use of this data.

22. For patient subgroups identified in section 2(d)(iii) who are poorly represented in the local context (e.g., Native American women in Boston), AI solution performance may have wide confidence bounds. Ensure that performance measures account for uncertainty.
23. Sometimes, removing SDOH and demographic data as model features may decrease model performance. For example, removing race worsened model performance (Khor S, Haupt EC, Hahn EE, Lyons LJL, Shankaran V, Bansal A. Racial and Ethnic Bias in Risk Prediction Models for Colorectal Cancer Recurrence When Race and Ethnicity Are Omitted as Predictors. JAMA Netw Open.

STAKEHOLDERS: S Strategic; O Operational; C Clinical; T Technical; R Regulatory; P Patient; CC Clinical Champion

EQUITY PRINCIPLES: Accountability; Fitness for purpose; Reliability and validity; Fairness; Transparency

SOURCES OF DATA: Local healthcare retrospective data; Local healthcare prospective data; Training data; Literature review; Local non-healthcare data; Qualitative data; Organizational data

FOR AN AI SOLUTION THAT IS BEING NEWLY DEVELOPED

For an AI solution that is being newly developed, skip to 5(j).

h **Assess prospective model performance and compare it to retrospective model performance across disadvantaged and advantaged patient subgroups.**

T

i. Conduct retrospective analysis of model performance on local healthcare retrospective data across disadvantaged and advantaged patient subgroups identified in 2(d)(iii) against fairness metrics identified in 3(b)(v).

ii. Conduct prospective analysis of model performance on local healthcare prospective data across disadvantaged and advantaged patient subgroups identified in 2(d)(iii) against fairness metrics identified in 3(b)(v).

iii. Ensure that model performance across disadvantaged and advantaged patient subgroups identified in 2(d)(iii) is consistent across local retrospective and local prospective settings against fairness metrics identified in 3(b)(v).

i **Perform comprehensive assessment of model performance across disadvantaged and advantaged patient subgroups.**

T

i. Conduct retrospective analysis of model performance on local healthcare retrospective data across disadvantaged and advantaged patient subgroups identified in 2(d)(iii) against fairness metrics identified in 3(b)(v).

ii. Conduct prospective analysis of model performance on local healthcare prospective data across disadvantaged and advantaged patient subgroups identified in 2(d)(iii) against fairness metrics identified in 3(b)(v).

iii. Compare model performance on retrospective data derived in 5(i)(i) and model performance on prospective data derived in 5(i)(ii) to model performance data gathered from external settings in 2(f)(i).

iv. Ensure that model performance across disadvantaged and advantaged patient subgroups identified in 2(d)(iii) is consistent across local retrospective, local prospective, and only if the AI product was built externally against fairness metrics identified in 3(b)(v).

j **If the model performs worse for a certain patient subgroup on local prospective healthcare data, consider adapting the model or its use to help minimize negative impacts on inequities.**

T

i. Consider the following to improve model performance for the patient subgroups negatively affected by poor model performance:
- Oversample patients from the affected subgroup to build the model.
- Source additional data and features that are specifically important for the affected patient subgroup.
- Process data differently, including different approaches to imputing data, that are specific to the affected patient subgroup.
- Build a separate model specifically targeted for the affected patient subgroup.
- Change the model threshold for the affected patient subgroup. Assess whether a different threshold could achieve similar levels of sensitivity, even if positive predictive value or precision suffers.[24]

ii. Consider the following if model performance for a certain patient subgroup cannot be improved:
- Narrow the scope of use of the model to only be used on patient subgroups for whom equity objectives defined in 3(a)(ii) are attainable. Note that it is possible that a model that performs poorly on disadvantaged patient subgroups can still achieve equity objectives. [25]
- Consider alternative, non-AI solutions (return to 2(a)).

24. In some cases (e.g., targeted vaccination campaigns and equity efforts in doulas), it may be appropriate to provide unequal resourcing and intervention to different groups of patients to close the potential inequities.
25. For example, imagine that there is a model built with an intention to de-escalate an intervention that currently harms Black patients at a higher rate than White patients. The model is found to be less accurate on Black patients than White patients. However, as long as the model identifies some Black patients, differences in model performance should not be too concerning because it is still better than the

STAKEHOLDERS: S Strategic; O Operational; C Clinical; T Technical; R Regulatory; P Patient; CC Clinical Champion

EQUITY PRINCIPLES: Accountability; Fitness for purpose; Reliability and validity; Fairness; Transparency

SOURCES OF DATA: Local healthcare retrospective data; Local healthcare prospective data; Training data; Literature review; Local non-healthcare data; Qualitative data; Organizational data

k **Ensure that model performance aligns with the equity objectives.**

T

i. Ensure that model performance measures derived in 5(h)(iii) or 5(i)(iv) align with the requirements for achieving the equity objectives defined in 3(a)(ii). [26]

ii. If model performance does not align with equity objectives defined in 3(a)(ii), consider either selecting an alternate solution (return to 2(a)) or adapting the model (return to 3(a)).

l **Conduct a prospective pilot study to validate whether the AI solution achieves equity objectives.**

C P T

i. Recruit a diverse sample of clinical end-users and patients. Ensure that disadvantaged patient subgroups identified in 2(d)(iii) are adequately represented in the pilot study context.

ii. Randomly assign units, frontline clinicians, or individual patients to control (baseline approach without the new AI solution) and intervention (with the new AI solution) groups. If unable to conduct a randomized study, conduct an observational study that controls for confounders, such as interrupted time series or differences-in-differences study designs.

iii. Couple quantitative assessment of patient outcomes with qualitative analyses of frontline clinicians and patients. Ensure that patient representatives from disadvantaged patient subgroups identified in 2(d)(iii) are included in qualitative research efforts.

iv. Assess performance of the model according to fairness metrics identified in 3(b)(iv) during the prospective pilot study.

v. Assess the degree of adoption and effectiveness of non-technical solution components defined in 4(c)(i).

vi. Assess measurably progress towards achieving equity objectives specified in 3(a)(ii).

vii. Report results of 5(l)(iv), 5(l)(v), and 5(l)(vi) to frontline clinicians and local patient community members.

26. For example, a model exhibiting minimal lead time (i.e., model identifies the outcome of interest shortly before the event occurs) would be ineffective at addressing a diagnostic inequity (i.e., more severe disease progression at the time of diagnosis).

STAKEHOLDERS: S Strategic; O Operational; C Clinical; T Technical; R Regulatory; P Patient; CC Clinical Champion

EQUITY PRINCIPLES: Accountability; Fitness for purpose; Reliability and validity; Fairness; Transparency

SOURCES OF DATA: Local healthcare retrospective data; Local healthcare prospective data; Training data; Literature review; Local non-healthcare data; Qualitative data; Organizational data

6. Execute AI solution roll out

Decision Point 6 describes how to disseminate information about AI products to affected clinicians, including end-users, and manage changes in the workflow caused by clinical integration of AI products. It also explains how to prevent misuse of AI products beyond their intended scope.

a

Document information about the model development and implementation and share it with clinical end-users, members of disadvantaged patient subgroups, and others who may be affected by use of the model.

C O P T

i. Create communication material with the following information:
   - Equity objectives defined in 3(a)(ii)
   - Fairness metrics identified in 3(b)(v)
   - Information about non-technical solution components defined in 4(c)(i) meant to support achieving the equity objectives
   - Model performance across disadvantaged patient subgroups identified in 5(h)(iii) or 5(i)(iv)
   - Implementation domain, directions, workflow, and warnings

ii. Tailor the communication materials to be understandable and relevant to clinical end-users, members of disadvantaged patient subgroups identified in 2(d)(iii), and others who may be affected by use of the model.

iii. Secure approval of communication materials for sharing.

iv. Develop a communication plan to share the materials with the personnel identified in 6(a)(ii).

v. Operationalize the communication plan and ensure the materials reach the personnel identified in 6(a)(ii).

b

Educate clinical end-users about potential bias in using the solution.

C

i. Make clinical end-users aware of baseline inequities surfaced in 2(d)(ii) and the equity objectives set in 3(a)(ii).

ii. Educate clinical end-users about confirmation bias and its harmful consequences on health inequity.

iii. Provide clinical end-users direct and explicit instructions on what they are supposed to do with the model output and any specific actions they need to take to achieve equity objectives set in 3(a)(ii).[27]

27 Meanwhile, it is important to respect clinician autonomy.

STAKEHOLDERS: S Strategic; O Operational; C Clinical; T Technical; R Regulatory; P Patient; CC Clinical Champion

EQUITY PRINCIPLES: Accountability; Fitness for purpose; Reliability and validity; Fairness; Transparency

SOURCES OF DATA: Local healthcare retrospective data; Local healthcare prospective data; Training data; Literature review; Local non-healthcare data; Qualitative data; Organizational data

c Where applicable, seek an approval from an institutional review board, ethical review board, or research ethics board to implement and use the AI solution in clinical practice.

R

d After rollout, continue to seek feedback from clinical end-users and members of disadvantaged and advantaged patient subgroups to achieve equity objectives.

C P T

i. Review recommendations surfaced in 4(a) and the non-technical solution components defined in 4(c)(i).

ii. Create a communication plan to seek feedback on the solution and surface unanticipated challenges associated with the solution from clinical end-users and members of disadvantaged and advantaged patient subgroups identified in 2(d)(iii).

iii. Seek feedback from clinical end-users and members of disadvantaged and advantaged patient subgroups identified in 2(d)(iii) on how well the solution progresses toward achieving equity objectives defined in 3(a)(ii).

iv. If challenges persist and limit progress towards achieving equity objectives, consider updating the AI solution or workflow (see Decision Point 8).

STAKEHOLDERS: S Strategic; O Operational; C Clinical; T Technical; R Regulatory; P Patient; CC Clinical Champion

EQUITY PRINCIPLES: Accountability; Fitness for purpose; Reliability and validity; Fairness; Transparency

SOURCES OF DATA: Local healthcare retrospective data; Local healthcare prospective data; Training data; Literature review; Local non-healthcare data; Qualitative data; Organizational data

7. Monitor AI Solution

Decision Point 7 focuses on monitoring AI solutions and affected workflow over time and sustaining improved outcomes achieved through AI use. It also highlights the importance of conducting audits of AI products that complement ongoing monitoring and proactively identifying risks not envisioned at the time of clinical integration.

a

Regularly monitor the model performance across disadvantaged and advantaged patient subgroups.

P T

i. Monitor model performance across disadvantaged and advantaged patient subgroups identified in 2(d)(iii).

ii. Monitor fairness metrics identified in 3(b)(iv).

iii. Share monitoring outcomes with clinical end-users, members of disadvantaged and advantaged patient subgroups identified in 2(d)(iii), and others who may be affected by use of the model via pre-established communication plan in 6(c)(ii).

b

Regularly monitor the work environment of the solution across disadvantaged and advantaged patient subgroups.

O P T

i. Monitor progress towards achieving equity objectives specified in 3(a)(ii).

ii. Monitor the degree of adoption of non-technical solution components defined in 4(c)(i).

iii. Share monitoring outcomes with clinical end-users, members of disadvantaged and advantaged patient subgroups identified in 2(d)(iii), and others who may be affected by use of the model through the pre-established communication plan.

iv. Seek feedback from clinical end-users, members of disadvantaged and advantaged patient subgroups identified in 2(d)(iii), and others who may be affected by use of the model to understand their experience with the AI solution and progress towards achieving equity objectives.

STAKEHOLDERS: S Strategic; O Operational; C Clinical; T Technical; R Regulatory; P Patient; CC Clinical Champion

EQUITY PRINCIPLES: Accountability; Fitness for purpose; Reliability and validity; Fairness; Transparency

SOURCES OF DATA: Local healthcare retrospective data; Local healthcare prospective data; Training data; Literature review; Local non-healthcare data; Qualitative data; Organizational data

C **Regularly monitor health inequities across disadvantaged and advantaged patient subgroups.**

C O P T

i. Calculate health inequities identified in 2(d)(ii) for disadvantaged and advantaged patient subgroups identified in 2(d)(iii).

ii. Compare the level of inequities calculated in 7(c)(i) with the baseline level calculated in 2(d)(ii).

iii. Ensure that any changes in the level of inequities align with equity objectives defined in 3(a)(ii).

iv. Consult with frontline clinicians, management, and members of disadvantaged patient subgroups identified in 2(d)(iii) on an ongoing basis and validate whether the quantitative measures reflect reality on the ground.

v. If inequities worsen in ways that were not anticipated and undermine equity objectives, proceed to 8(a) and 8(b).[28]

28 Note that failing to meet equity objectives may be due to ineffective implementation or adoption of non- technical solution components defined in 4(c)(i) rather than inequities in AI model performance.

STAKEHOLDERS: S Strategic; O Operational; C Clinical; T Technical; R Regulatory; P Patient; CC Clinical Champion

EQUITY PRINCIPLES: Accountability; Fitness for purpose; Reliability and validity; Fairness; Transparency

SOURCES OF DATA: Local healthcare retrospective data; Local healthcare prospective data; Training data; Literature review; Local non-healthcare data; Qualitative data; Organizational data

8. Update or decommission the AI solution

Decision Point 8 focuses on updating or decommissioning AI products, adapting work environments affected by AI use, and expanding the use of AI products to new settings. It also describes how to minimize disruptions and eliminate harms that result from decommissioning AI products and disseminating information about AI product updates to affected clinicians.

a **Determine updates to the model or its work environment.**

C O P T

i. If metrics monitored in 7(a)(i) and 7(a)(ii) reveal changes, engage a cross-functional team to determine the need for updates.

ii. When changes in model performance across disadvantaged patient subgroups identified in 2(d)(iii) or changes in fairness metrics identified in 3(b)(iv) are observed, consider executing activities described in 5(j)(i) to improve model performance for the disadvantaged patient subgroups.

iii. When changes in the effectiveness of non-technical solution components defined in 4(c)(i) or changes in progress toward achieving equity objectives defined in 3(a)(ii) are observed, reassess model performance as well as non-technical solution components defined in 4(c)(i).[29] Consider revisiting 4(a) and 4(c) to update non-technical interventions that can better achieve equity objectives.

iv. When changes in the model or workflow are made, ensure to communicate information about the changes to all who may be affected by the changes through the pre-established communication plan in 6(c)(ii).

b **If updating the model or its work environment does not improve model performance or fails to improve progress towards equity objectives, consider decommissioning the model.**

C O P

i. Revisit potential solutions identified in 2(a) and consider alternative solutions that are well equipped to improve health equity for disadvantaged patient subgroups identified in 2(d)(iii).

ii. Conduct a final assessment of the model and report outcomes to both frontline clinicians and members of disadvantaged patient subgroups identified in 2(d)(iii).

iii. Consider decommissioning the model, if an alternative solution is better equipped to improve health equity for disadvantaged patient subgroups or the results of the final assessment are poor based on decommissioning cut-offs established in 3(b)(iv) and 3(a)(ii).

iv. When a decision to decommission the model is made, ensure to communicate information about the decision to all who may be affected by the decision through the pre-established communication plan in 6(c)(ii).

29. For example, there may be insufficient staffing or capacity available to respond to AI model outputs to address inequities.

STAKEHOLDERS: S Strategic; O Operational; C Clinical; T Technical; R Regulatory; P Patient; CC Clinical Champion

EQUITY PRINCIPLES: Accountability; Fitness for purpose; Reliability and validity; Fairness; Transparency

SOURCES OF DATA: Local healthcare retrospective data; Local healthcare prospective data; Training data; Literature review; Local non-healthcare data; Qualitative data; Organizational data

C **If the model successfully achieved equity objectives and there is interest to expand model use, evaluate appropriate technical and non-technical resources to expand the AI solution to new settings.**

C O P

i. Assess baseline health inequities in the new implementation context, following a procedure described in 2(d)(ii).

ii. Assess model performance across disadvantaged patient subgroups identified in 2(d)(iii) in the new implementation context.

iii. Define equity objectives for the new implementation context, following a procedure described in 3(a)(ii).

iv. Design non-technical solution components to help achieve equity objectives following procedures described in 4(c) in the new implementation context.

v. Assess model performance following procedures described in 5(f) in the new implementation context.

vi. Conduct a prospective pilot study following procedures described in 5(l) in the new implementation context.

Glossary

Actual label
A.k.a. actual target. Downstream outcome used to train or evaluate a model.

Advantaged patient group
A group of patients who are least likely to be negatively impacted by health inequities.

Baseline population
Population in a geographic setting where a healthcare delivery organization is based.

Computable ideal label
A computable proxy for an ideal label that can be derived from structured or unstructured data elements available within local data.

Confirmation bias
The tendency to gather evidence that confirms preexisting expectations, typically by emphasizing or pursuing supporting evidence, while dismissing or failing to seek contradictory evidence.

Cross-functional team
A team composed of personnel with a variety of expertise who come together to achieve a common goal. In this framework, a cross-functional team is composed of all active stakeholders listed in each procedure.

Disadvantaged patient group
A group of patients who are most likely to be negatively impacted by health inequities.

End-user
A person who uses and interacts with the product. Typically, end-users are frontline clinicians.

Feature
Data in its final form post cleaning and QA that is being fed directly into model training and evaluation or for model inference.

Hidden stratification
The data contains unrecognized subsets of cases which may affect model training, measured model performance, and most importantly the clinical outcomes.

Ideal label
A.k.a. ideal target or ground truth outcome. Clinical outcome used for model training that is validated by clinicians when pristine and complete information about patients is available for model development.

Label choice bias
A mismatch between the ideal label and the actual label that results in worse model performance for a disadvantaged patient group.

Local healthcare retrospective data
Historical healthcare data that is curated within the primary healthcare delivery organization seeking to adopt an AI product. The local data can be sourced from a variety of systems, including the EHR, radiology PACS system, medical claims, audit logs, electrocardiograms, and high-frequency vital sign monitors. When a model is internally developed, the local healthcare retrospective data set is used for training the model.

Local healthcare prospective data
Real-time healthcare data that is curated within the primary healthcare delivery organization seeking to adopt an AI product. The local data can be sourced from a variety of systems, including the EHR, radiology PACS system, medical claims, audit logs, electrocardiograms, and high-frequency vital sign monitors. The local healthcare prospective data set is used for validating a model during a ‘silent trial’ and for using the model in clinical care.

Local non-healthcare data
Non-healthcare data that is curated within a geographic setting where a healthcare delivery organization is based. The local non-healthcare data can be derived from a variety of external sources, including US Census.

Missing indicator
A binary variable that indicates whether a model feature is missing or not.

Patient group
A collection of patients with similar experiences, cultural norms, practices, or way of life.

Representation bias
A mismatch between the demographic makeup of the local population, especially regarding disadvantaged patient groups, and the demographic makeup of the training data.

Training data
Data used for training a model. When the model is externally developed, the training data set contains data from an external source. When the model is internally developed, the training data set is sourced from local healthcare retrospective data.

Work environment
A sociotechnical environment where an AI solution is being used and interacted by its end-users and others who are affected by its use

Copyright