ASPE | OFFICE OF HEALTH POLICY

OS PCOR TF | OFFICE OF THE SECRETARY PATIENT-CENTERED OUTCOMES RESEARCH TRUST FUND

REPORT

Trustworthy Artificial Intelligence (TAI) for Patient-Centered Outcomes Research (PCOR)

Prepared for
The Office of the Assistant Secretary for Planning and Evaluation (ASPE)
at the U.S. Department of Health and Human Services

by
NORC at the University of Chicago

September 2023

Office Of The Assistant Secretary For Planning And Evaluation

The Assistant Secretary for Planning and Evaluation (ASPE) advises the Secretary of the U.S. Department of Health and Human Services (HHS) on policy development in health, disability, human services, data, and science; and provides advice and analysis on economic policy. ASPE leads special initiatives; coordinates the Department's evaluation, research, and demonstration activities; and manages cross-Department planning activities such as strategic planning, legislative planning, and review of regulations. Integral to this role, ASPE conducts research and evaluation studies; develops policy analyses; and estimates the cost and benefits of policy alternatives under consideration by the Department or Congress.

The Office Of Health Policy

The Office of Health Policy (HP) provides a cross-cutting policy perspective that bridges Departmental programs, public and private sector activities, and the research community, in order to develop, analyze, coordinate and provide leadership on health policy issues for the Secretary. HP carries out this mission by conducting policy, economic and budget analyses, assisting in the development and review of regulations, assisting in the development and formulation of budgets and legislation, and assisting in survey design efforts, as well as conducting and coordinating research, evaluation, and information dissemination on issues relating to health policy.

Office Of The Secretary – Patient-Centered Outcomes Research Trust Fund

The Office of the Secretary Patient-Centered Outcomes Research Trust Fund (OS-PCORTF) was established as part of the 2010 Patient Protection and Affordable Care Act and is charged to build data capacity for patient-centered outcomes research. Coordinated by ASPE on behalf of the Department, OS-PCORTF has funded a rich portfolio of projects to meet emerging U.S. Department of Health and Human Services policy priorities and fill gaps in data infrastructure to enhance capabilities to collect, link, and analyze data for patient-centered outcomes research. For more information, visit https://aspe.hhs.gov/collaborations-committees-advisory-groups/os-pcortf

This report was funded by the Office of the Secretary Patient-Centered Outcomes Research Trust Fund (OS-PCORTF) under Contract Number HHSP233201500048II of the HHS Office of the Assistant Secretary for Planning and Evaluation (ASPE). The work was carried out by NORC at the University of Chicago and ASPE. The authors are solely responsible for this document's contents, findings, and conclusions, which do not necessarily represent the views of HHS, ASPE, or NORC. Readers should not interpret any statement in this product as an official position of ASPE or of HHS.

Suggested Citation: Dullabh P, Dhopeshwarkar R, Leaphart D, Peterson C, Gauthreaux N, Grigorescu V, Eloso J, Wei S. Trustworthy Artificial Intelligence (TAI) for Patient-Centered Outcomes Research (PCOR). Office of the Assistant Secretary for Planning and Evaluation, U.S. Department of Health and Human Services. September 2023.

Contributing Authors

Prashila Dullabh,* MD, Vice President and Senior Fellow and Rina Dhopeshwarkar,* MPH, Principal Research Scientist, NORC

Desirae Leaphart, MPH, Research Scientist, NORC

Caroline Peterson, MPH, Senior Research Associate II, NORC

Nicole Gauthreaux, MPH, Senior Research Associate I, NORC

Violanda Grigorescu, MD, MSPH, Senior Health Scientist, ASPE

Jessica Eloso, PharmD, BCPS, CDC Public Health Informatics Fellow, ASPE

Sara Wei, MHA, Public Health Analyst, ASPE

*These authors contributed equally to the development of this report

Subject Matter Experts

Michael E. Matheny, MD, MS, MPH, FACMI, Professor, Vanderbilt University Medical Center

Maia Hightower, MD, MPH, MBA, Chief Medical Information Officer, University of Utah Health

Key Informants

Brian Anderson, MD, Chief Digital Health Physician at MITRE

Christine Dymek, EdD, Director of the Digital Healthcare Research Division, Center for Evidence and Practice Improvement (CEPI), Agency for Healthcare Research and Quality (AHRQ)

Susan Gregurick, PhD, Associate Director for Data Science (ADDS) and Director of the Office of Data Science Strategy (ODSS), National Institutes of Health (NIH)

Travis Hoppe, PhD, Associate Director for Data Analytics and Data Science, National Center for Health Statistics (NCHS), Centers for Disease Control and Prevention (CDC)

Scott Lee, PhD, Statistician, Centers for Disease Control and Prevention (CDC)

Keith Marsolo, PhD, Associate Professor, Department of Population Health Sciences, Duke University School of Medicine

Michael Morgan, Project Manager, Office of Regulatory Science and Innovation (ORSI), Food and Drug Administration (FDA)

Tina Morrison, PhD, Director of the Office of Regulatory Science and Innovation (ORSI), Food and Drug Administration (FDA)

Eliel Oliveira, MS, MBA, Director of the Research Data Infrastructure, Department of Population Health, Dell Medical School of the University of Texas at Austin

Papia Paul, MS, MPA, Public Health Analyst, Office of the National Coordinator for Health Information Technology (ONC)

Christina Silcox, PhD, Research Director for Digital Health, Duke-Margolis Center for Health Policy

Jeffery Smith, MPP, Deputy Director, Certification & Testing Division, Office of the National Coordinator for Health Information Technology (ONC)

Adam Wong, MPP, Senior Innovation Analyst, Office of the National Coordinator for Health Information Technology (ONC)

Table of Contents

Table of Exhibits

Executive Summary

Background. The rise of artificial intelligence (AI) in health care and health care research has stimulated discourse on AI's trustworthiness and its potential to cause harm. To provide guidance for U.S. Department of Health and Human Services (HHS) agencies on how to manage AI at all stages of the technology's lifecycle, the Office of the Chief AI Officer published the Trustworthy AI (TAI) Playbook in September 2021. The Playbook defined TAI as the "design, development, acquisition, and use of AI in a manner that fosters public trust and confidence while protecting privacy, civil rights, civil liberties, and American values, consistent with applicable laws." The Playbook also outlined six TAI principles: 1) fair/impartial, 2) transparent/explainable, 3) responsible/accountable, 4) robust/reliable, 5) privacy, and 6) safe/secure.

Considerations for implementing TAI principles for health care research, and patient-centered outcomes research (PCOR) in particular, are not explicitly addressed in the Playbook. Such considerations would be especially helpful for PCOR that the Office of the Secretary Patient-Centered Outcomes Research Trust Fund (OS-PCORTF) supports through data infrastructure capacity building.

Project Purpose. One goal of the 2020–2029 OS-PCORTF Strategic Plan is "to leverage leading technology solutions to improve data capacity for patient-centered outcomes and comparative clinical effectiveness research." Leading technology solutions include use of AI tools and methods. Under this goal, the HHS Office of the Assistant Secretary for Planning and Evaluation (ASPE) commissioned NORC at the University of Chicago (NORC) to develop a report to inform a better understanding about how the Playbook's TAI principles can be applied to OS-PCORTF projects. The TAI Playbook is a foundation to describe and understand the TAI principles. We used an adapted version of the National Library of Medicine (NLM) research lifecycle to highlight the connection between TAI principles and PCOR.

Methods. We used two approaches to gather the information summarized in this report: 1) an environmental scan of gray and peer-reviewed literature; and 2) eleven key informant discussions with federal and non-federal stakeholder experts in AI who validated findings synthesized from the environmental scan. We adapted the NLM's research lifecycle and mapped the findings to the six phases of the lifecycle to contextualize the considerations for the OS-PCORTF community and PCOR researchers.

Results. Our findings are organized into three categories to inform how OS-PCORTF projects can adhere to the HHS Playbook's six TAI principles:

- ***Key Informant Reflections on Implementing the Six HHS TAI Principles***

  Key informant discussions noted that the six HHS TAI principles cover all salient ethical areas for consideration when using AI in PCOR, yet some principles are more difficult to implement and interpret than others. There was consensus that the privacy principle is the most intuitive and easiest to implement. Key informants described the transparent/explainable principle as difficult to implement for "black box" AI models, where the decision-making process may not be explained. Nearly all key informants agreed that the fair/impartial principle is the most difficult to conceptualize and to implement. Key informants also reacted to the Playbook's definition of the safe/secure principle, noting that there should be more emphasis on protecting the safety and security of individuals from harm that may result from use of AI in research.

- ***Considerations for the OS-PCORTF Community and PCOR Researchers in Adhering to TAI Principles***

  This report describes 15 considerations for OS-PCORTF project adherence to the six TAI principles. We identified these considerations through the environmental scan of gray and peer-reviewed literature, refined and validated them through key informant discussions, and organized them by each of the six research lifecycle phases. Considerations that apply to all phases include ensuring patient privacy and safety are protected, evaluating tradeoffs between principles, and iteratively examining principles in every phase. We also identified TAI considerations for researchers specific to each research lifecycle phase. When planning a research project, considerations include determining the use case for the AI algorithm and establishing proper structures and procedures. During data acquisition, researchers can consider determining the appropriate volume, quality, and representativeness of the data. When preparing data for AI, considerations include augmenting the data and reducing errors that occurred during data collection. When analyzing data and maintaining AI models, researchers should test and evaluate models continuously for performance and for risk of bias or adverse events. Finally, when sharing results or reusing the AI algorithm, researchers can consider promoting transparency in their reporting.

- ***Opportunities for the OS-PCORTF to Support Work that Promotes Adherence to TAI Principles***

  To adhere to the HHS TAI principles, OS-PCORTF may consider 14 opportunities to support improvements to tools, resources, and methods/techniques. We used the environmental scan of gray and peer-reviewed literature to identify the opportunities and refined them through key informant discussions with federal and non-federal informants involved in AI-enabled research. We identified one opportunity related to governance, which describes updating documents that the OS-PCORTF has produced to address ethical considerations around using AI algorithms or methods. Five identified opportunities related to data, including development and use of standardized data sets, methods to augment training data, synthetic data modules, federated data models, and foundation models. Finally, we identified eight opportunities related to developing tools and resources, such as implementation guides, evaluation methods and metrics guidance, curated repositories of tools addressing bias and transparency, inventories of AI-related efforts, core resources for researchers using AI, and forums to discuss tools and resources.

Conclusion. Our findings highlight that the TAI principles outlined in the HHS TAI Playbook are important to implement when using AI for PCOR, but that implementation is complex and use case dependent. As the OS-PCORTF portfolio expands its projects that leverage AI, our report can be used as resource by the OS-PCORTF community, policymakers, PCOR researchers and others to inform application of the HHS TAI principles.

1. Introduction

The rapid evolution and expansion of artificial intelligence (AI) has led researchers to develop dozens of guidelines, frameworks, and tools to foster ethical AI. However, there is little concrete guidance on how to implement ethics principles in practice.[1] This gap in guidance, along with the inherently subjective interpretation of ethics principles, poses a challenge for organizations trying to implement and assess trustworthy AI (TAI).[2]

In March 2021, the U.S. Department of Health and Human Services (HHS) established the Office of the Chief AI Officer (OCAIO) to create processes supporting TAI development across HHS agencies. In 2021, the HHS OCAIO published a TAI Playbook outlining six core TAI principles and identifying actions to advance TAI for different types of AI solutions.[3]

The federal definition of TAI is the "design, development, acquisition, and use of AI in a manner that fosters public trust and confidence while protecting privacy, civil rights, civil liberties, and American values, consistent with applicable laws."[4] Although a large body of literature has been developed on AI ethics in *health care delivery*, far fewer publications focus on AI ethics in *health care research*, particularly patient-centered outcomes research (PCOR).[5] PCOR's focus on helping patients and caregivers "communicate and make informed health care decisions, allowing their voices to be heard in assessing the value of health care options" makes it critical for PCOR researchers using AI methods to understand and apply TAI principles throughout their research.[6]

Effective application of TAI principles in PCOR is relevant to the HHS Office of the Assistant Secretary for Planning and Evaluation (ASPE). Under delegation of authority by the Secretary of HHS and through the administration of the Office of the Secretary Patient-Centered Outcomes Research Trust Fund (OS-PCORTF), ASPE coordinates across relevant federal health programs to build data capacity for PCOR. The OS-PCORTF's strategic vision is "better data for patient-centered outcomes research to improve evidence generation, decision-making, and health outcomes for all Americans."[7]

Artificial intelligence (AI) "enables computer systems to perform tasks normally requiring human intelligence."[8]

Machine learning (ML), a type of AI, is "the use and development of computer systems that are able to learn and adapt without following explicit instructions, by using algorithms and statistical models to analyze and draw inferences from patterns in data."[9]

Predictive AI includes ML, statistical modeling, and data mining techniques that can support predictive analytics.[10]

The OS-PCORTF's Strategic Plan (2020 – 2029)[11] charts a course for strengthening data capacity for PCOR. The Plan's third goal is to leverage advanced technology solutions to improve the use of large volumes of data, as well as the variety and timeliness of data available for PCOR.[11] Such technology solutions include AI tools and machine learning (ML) techniques, which several OS-PCORTF-funded projects explore and use to improve the richness and robustness of evidence generated.[12, 13, 14] It is critical for ASPE to identify ways to ensure that future OS-PCORTF work using AI abides by TAI principles.

This report presents the findings of an environmental scan and key informant discussions conducted to better understand how TAI principles can be applied in the use of predictive AI in OS-PCORTF and PCOR projects. In the report, we do not address generative AI, which consists of deep learning models that can generate text, images, data, and other content.[15]

2. Background

In recent years, AI has become an important tool in precision medicine and biomedical research to leverage and analyze growing volumes of health care data, with important early successes in medical imaging research.[16] There are a variety of AI methods used for health care data, including ML and natural language processing (NLP).[3] AI can be used to mine and analyze large-scale data repositories—including data from electronic health record (EHR) systems, claims data, clinical registries, and genomics data—and to interpret outputs for clinical decision-making and population health.[17, 18] In addition, AI can be valuable for conducting clinical trials[19] and for engaging patient stakeholders in research.[20] Although not the focus of this report, generative AI tools such as ChatGPT can be used to improve efficiency in research, for example, by assisting in identifying relevant literature or code or scripts for ML models. We expect that as the field of AI evolves, applications of AI in health care research will continue to expand.

Illustrative applications of AI in health care research:[21]

- Developing matched cohorts for clinical trials, especially in oncology research.[22]
- Developing health information technology tools to aid in preventing medication errors and to improve patient safety.[23]
- Collecting health data using conversational agents, such as chatbots.[24]

The rise of AI in health care and health care research has sparked discussion about the trustworthiness of AI and its potential to cause harm, whether through breaches of patient privacy or by delivering systematically biased results. Researchers, clinical professional societies, and government agencies have all acknowledged that the widespread and increasing use of AI in health care and health care research may perpetuate inequities and needs oversight.[25] Increasingly, federal agencies and other organizations (listed below) are developing tools and guidance to mitigate potential harm due to AI by promoting TAI:

- The White House has made responsible AI research, development, and deployment a priority in the national agenda and has developed resources, such as the Blueprint for an AI Bill of Rights, to manage risks to national security.[26]

- The National Institute of Standards and Technology (NIST) developed an AI Risk Management Framework (AI RMF 1.0) in 2023 to help individuals and organizations better manage risks associated with AI; NIST has also released a playbook, roadmap, crosswalk, and explainer video to supplement the framework.[27]

- The Coalition for Health AI (CHAI), led by the MITRE Corporation, aims to develop "guidelines and guardrails" through a consensus-driven framework for health AI systems; CHAI has developed a draft blueprint for TAI implementation guidance.[28]

- The National Academy of Medicine (NAM) has initiated the Artificial Intelligence Code of Conduct project, which aims to provide a guiding framework to ensure AI algorithms used in health care and health care research perform "accurately, safely, reliably, and ethically in the service of better health for all." The project involves national multidisciplinary leaders in its efforts to advance TAI.[29]

3. Overview of HHS TAI Principles

The HHS TAI Playbook published in 2021 supports TAI development across HHS, outlining six core TAI principles and help to identify actions to advance TAI for different types of AI solutions.[3] Exhibit 1 lists

the definition for each principle, together with potential consequences of not aligning with the principles.

Exhibit 1. The Six TAI Principles from the HHS TAI Playbook and Potential Consequences of Nonalignment[3]

TAI Principle and Description[3]	Consequences of Nonalignment[30]
Fair/Impartial AI applications should include checks from internal and external stakeholders to help ensure equitable application across all participants.	Algorithms based on data that are inherently biased can result in research conclusions that perpetuate health inequities and that produce or reinforce negative health outcomes that disproportionately impact one group over another.
Transparent/Explainable All relevant individuals should understand how their data is being used and how AI systems make decisions; algorithms, attributes, and correlations should be open to inspection.	Lack of transparency can result in algorithmic systems that are hard to control, monitor, and correct (that is, the "black box" issue) and will result in lack of trust from key stakeholders and the public.
Responsible/Accountable Policies should outline governance and who is held responsible for all aspects of the AI solution (for example, initiation, development, outputs, decommissioning).	If responsibility for algorithmic systems is unclear, and if harm results from use of the algorithms, it will be difficult to know who to hold responsible for addressing and preventing further harm.
Robust/Reliable AI systems should have the ability to learn from humans and other systems and produce accurate and reliable outputs consistent with the original design.	Algorithms that are unreliable and/or inaccurate have a higher chance of producing research conclusions that are incorrect, which may harm patients and result in negative health outcomes, further eroding stakeholder and public trust.
Privacy The privacy of individuals, groups, or entities should be respected, and their data should not be used beyond its intended and stated use; data used has been approved by the data owner or steward.	If patients feel that their privacy was violated, they are unlikely to participate in research and may mistrust the health care system.
Safe/Secure AI systems should be protected from risks (including cyber) that may directly or indirectly cause physical and/or digital harm to any individual, group, or entity.	If access to protected patient information is compromised, information may be exploited by unauthorized entities; as a result, the organization using the AI system may lose credibility.

4. Report Purpose

The HHS TAI Playbook supports leaders across HHS in applying TAI principles for developing and deploying AI solutions; however, the principles must be described in the context of research, specifically PCOR. This report presents key considerations on how the TAI principles can be applied in the context of OS-PCORTF projects and PCOR more broadly. We use the TAI Playbook as a foundation, mapping to an adapted National Library of Medicine (NLM) research lifecycle construct to offer context for the principles. The report focuses on how to apply TAI principles when using predictive AI models for PCOR.

5. Methods

We gathered information for this report through: 1) an environmental scan of gray and peer-reviewed literature, and 2) key informant discussions with AI and PCOR experts to validate findings from the environmental scan.

Environmental scan. We used three overarching questions to guide our environmental scan:

1. What are the special considerations for applying TAI principles for OS-PCORTF/PCOR projects?
2. What strategies have been used to review TAI principles within AI solutions for improving PCOR data infrastructure and PCOR more broadly?
3. How have other agencies and research organizations applied and used trustworthy principles in their AI-focused PCOR/health care research work?

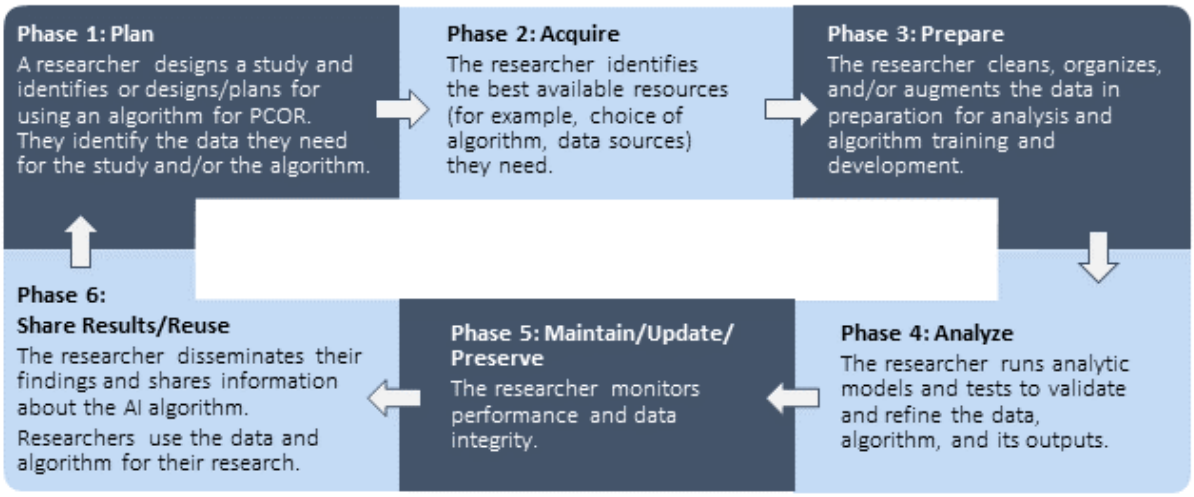
First, we used pre-specified search terms to search PubMed for peer-reviewed articles related to AI in PCOR and health care research more broadly, as well as to the six principles; see Exhibit A1 in Appendix A for search terms and Exhibit A3 for specific search strings. Next, we screened articles using a two-step process that involved a title/abstract review, followed by a full text review of the articles retained according to pre-specified inclusion/exclusion criteria; see Exhibit A2 in Appendix A. We then conducted supplemental targeted searches for topics or specific concepts recommended to us by subject matter experts (SMEs) or key informants. Finally, we searched backward through the reference list of selected articles to add any additional relevant articles.

Exhibit A4 in Appendix A shows the article selection process. The initial PubMed search resulted in 331 articles for review, of which we removed 108 duplicates. We included an additional 63 resources at this step per recommendation of subject matter experts (for example, peer-reviewed articles, reports, tools, organizational updates on websites, and blogs from trusted sources). A total of 286 articles underwent the titles/abstracts review step. Upon applying our inclusion/exclusion criteria on the title/abstract, we identified a total of 170 articles for the full text review, including 34 articles we identified through backwards searching of references. Lastly, after reviewing the full text of the 170 articles, a final total of 132 articles were included.

Key informant discussions. We conducted 11 virtual, one-hour semi-structured key informant discussions with 13 experts in the field of AI, including individuals from federal agencies, academic research centers, and the private sector. The SMEs and key informants commented on our initial environmental scan findings, and we incorporated their feedback into our final synthesis. See Appendix B for the detailed protocol used to guide each discussion.

Analytic approach. We organized and synthesized findings by mapping considerations relevant for PCOR researchers to six distinct phases of the NLM research lifecycle construct that were applicable to an AI-enabled research lifecycle; see Exhibit 2.[31]

Exhibit 2. Six Phases of the AI-Enabled Research Lifecycle



6. Findings

We organized the findings of the report into three categories: 1) key informant reflections on implementing the TAI principles; 2) important considerations for PCOR researchers and the OS-PCORTF community to take into account when using the HHS TAI principles; and 3) potential opportunities for the OS-PCORTF to support project alignment with TAI principles.

6.1 Key Informant Reflections on Implementing the Six HHS TAI Principles

The TAI Playbook emphasizes the importance of all six principles to ensure TAI; however, the Playbook also acknowledges the challenges and tradeoffs related to implementing each principle. Key informants remarked that although TAI principles cover all the salient ethical areas, fully implementing each principle in a given project is typically a challenge.

Key informants noted that the **privacy** principle is the most intuitive and easiest to implement. According to both the literature and key informants, there are established, vetted tools and methods for protecting patient privacy in health care research already in use.[27, 32, 33] The tools and methods can be used for maintaining the privacy of data used in research that leverages AI. Key informants described the **transparent/explainable** principle as difficult to implement for black box AI models, where the decision-making process cannot always be explained. However, researchers can facilitate transparency by using tools and resources to document how the AI model was created; documentation should describe and characterize data sources, algorithm and parameter choices in model development and summarize performance validation in a way that a general audience can understand.

Nearly all key informants agreed that the **fair/impartial** principle is the most difficult to conceptualize and to implement. Implementing fairness in AI requires addressing the human factors that introduce bias in data and mitigating bias within existing data sets used for algorithm development. If there is underlying bias in the data used to train predictive AI algorithms, the resulting conclusions or predictions may further exacerbate existing inequities.[25] Several key informants also described a general lack of guidance regarding the definition of fairness and how to apply measurements for fairness. A lack of guidance stems from the inherent difficulty in defining and measuring fairness, as perceptions of fairness may vary with different cultural and institutional contexts.

> "'Fairness and impartiality' would be the most difficult because it is not intrinsic to building models… the idea of fairness and impartiality [is] not intrinsic or intuitive because society is complex, therefore the environment where models are deployed will be complex."
>
> **-Federal Key Informant**

Key informants reacted to the TAI Playbook's definition of the **safe/secure** principle, noting that there should be more emphasis on protecting the safety of individuals from potential harm caused by AI use,[34, 35] rather than focusing on the security of the AI system itself (for example, from malicious attacks).

Additionally, key informants noted there are tradeoffs with implementing the TAI principles.[36, 37, 38] The TAI Playbook emphasizes that "TAI principles are not mutually exclusive, and tradeoffs often exist when applying them."[3] Often a focus on one principle may require less adherence to another principle.[37] For example, when developing an AI-enabled health care tool, researchers often must select a cutoff point for action.[39] Selection of this cutoff point requires researchers to weigh maximizing *sensitivity*—identifying patients who would benefit from an intervention, aligning with the **fair/impartial** principle – against maximizing *specificity* – ensuring patients are not unnecessarily placed at risk by the intervention, aligning with the **safe/secure** principle.[39]

Recognizing that health care research is at an inflection point regarding the use of AI, informants noted that it is not feasible to stop AI use in the health care research context. Rather researchers must remain vigilant in building trust in AI solutions by being transparent about strengths, weaknesses, and limitations.

6.2 Considerations for OS-PCORTF Projects in Adhering to TAI Principles Across the Research Lifecycle

Below, we describe considerations that the OS-PCORTF community and PCOR researchers should take into account to align with the TAI principles; see Exhibit 3 for a summary of the considerations. We have organized the findings by the NLM's six research lifecycle phases (as shown in Exhibit 2), plus an initial overarching category that applies throughout the research lifecycle. Each consideration is tagged with the TAI principle(s) it addresses.

Exhibit 3. Considerations to Adhere to TAI Principles Across the Six Research Lifecycle Phases

Overarching Considerations
• **<u>Consideration 1:</u> Develop clear and effective protocols for data management and stewardship to ensure privacy and security are protected when developing, training, validating, and implementing AI models. (Transparent/Explainable; Responsible/Accountable; Privacy; Safe/Secure)**

Exhibit 3. Considerations to Adhere to TAI Principles Across the Six Research Lifecycle Phases
• Consideration 2: Consider the tradeoffs involved in taking action to improve one or more TAI principles within a project, and document decisions regarding tradeoffs throughout the research lifecycle. (Fair/Impartial; Transparent/Explainable; Responsible/Accountable; Robust/Reliable; Privacy; Safe/Secure)
• Consideration 3: Address and reassess AI solutions for trustworthiness in every phase of the research lifecycle. (Fair/Impartial; Transparent/Explainable; Responsible/Accountable; Robust/Reliable; Privacy; Safe/Secure)
Phase 1: Plan
• Consideration 4: Determine whether AI is appropriate for the research questions you are trying to answer, before beginning the project. (Fair/Impartial; Responsible/Accountable)
• Consideration 5: Develop a Steering Committee or Technical Expert Panel when leveraging AI solutions. The Committee or Panel should comprise experts in AI, data management, and information technology, as well as representatives of the patients and/or communities affected by the work. Where possible, consider the inclusion of experts that are also members of the affected communities. (Fair/Impartial; Transparent/Explainable; Responsible/Accountable; Robust/Reliable; Privacy; Safe/Secure)
• Consideration 6. Enhance transparency and protect privacy by implementing clear, thorough, data consent procedures that explicitly address use of patient data in AI models. (Transparent/Explainable; Responsible/Accountable; Privacy)
Phase 2: Acquire
• Consideration 7: Determine the appropriate volume and quality of data to support the identified problem and AI application, when identifying data sets. (Fair/Impartial; Robust/Reliable)
• Consideration 8: Assess whether the selected data sets represent the population being studied, when leveraging secondary data sets such as clinical registries, claims, or EHR data. (Fair/Impartial)
• Consideration 9: Explore the use of multiple, diverse, and high-quality data sources that support the identified use case for the AI model, including validation and training data sets. (Fair/Impartial; Transparent/Explainable; Responsible/Accountable; Robust/Reliable; Privacy)
Phase 3: Prepare
• Consideration 10: Consider techniques and methods that augment the data used when leveraging AI, especially in situations where multiple, diverse data sets cannot be acquired, or the amount of data needs to be artificially increased. (Fair/Impartial; Transparent/Explainable; Robust/Reliable; Privacy)
• Consideration 11: Consider instituting processes and protocols to reduce measurement error, missing data, and selection bias, any or all of which may occur during data collection. (Fair/Impartial; Robust/Reliable)

Exhibit 3. Considerations to Adhere to TAI Principles Across the Six Research Lifecycle Phases
Phase 4: Analyze
• **Consideration 12: Evaluate and test models for model performance, efficacy, accuracy, and adherence to principles using specific metrics and continue to monitor performance over time. (Fair/Impartial; Transparent/Explainable; Responsible/Accountable; Robust/Reliable;)**
• **Consideration 13: Test and assess algorithms and their outcomes for risk of bias using appropriate analytic tools and techniques, when using AI. (Fair/Impartial)**
Phase 5: Maintain/Update/Preserve
• **Consideration 14: Monitor and maintain deployed systems continuously to identify and address risks and adverse events, when leveraging AI algorithms. (Privacy; Safe/Secure)**
Phase 6: Share Results/Reuse
• **Consideration 15: Promote transparency by reporting comprehensively on the functionality, strengths, and weaknesses of an AI tool. (Transparent/Explainable)**

Overarching Considerations for OS-PCORTF Projects

There are three overarching TAI considerations that PCOR researchers should keep in mind throughout the research lifecycle:

Consideration 1. Develop clear and effective protocols for data management and stewardship to ensure privacy and security are protected when developing, training, validating, and implementing AI models. (Transparent/Explainable; Responsible/Accountable; Privacy; Safe/Secure)

Given the sensitivity of health care data and the high volume of data analyzed in many PCOR studies, privacy is a major consideration when implementing AI. Several key informants emphasized that protecting the privacy and security of individuals, groups, or entities is a fundamental concern in AI-enabled research, especially given that privacy breaches could significantly undermine efforts to promote trust in AI.

National Institute of Standards and Technology (NIST) Privacy Framework

- The NIST Privacy Framework is a tool for researchers and organizations to identify and manage privacy risks early in the development process; it includes a resource repository and roadmap to support implementation.[44]

Developing clear and effective data governance and stewardship protocols and data protection plans is key to ensuring the safety and security of AI. Data governance comprises organizational enterprise assets, policies, and practices informing "what data can be shared, with whom, under what conditions, and for what purposes."[40, 41] Effective governance requires designating entity(ies) responsible for all aspects of the AI lifecycle, and provides structures, systems, processes, and teams to help organizations develop a culture of risk management.[27] Experts in AI, data management, and information technology should be engaged in the development of governance frameworks.[42] Two ways that experts can be involved in governance is through the creation of steering committees (see Consideration 5) and through rigorous AI algorithm peer review or audit processes.

In addition, researchers should consider appropriate tools and methods to ensure and enhance patient privacy within the data management and stewardship protocols. Several key informants identified the need for privacy-enhancing technologies (PETs) for AI, as well as data breach minimizing methods such

as de-identification and aggregation for certain model outputs.[43, 44] The OS-PCORTF has dedicated efforts in this area by supporting research to evaluate privacy-preserving record linkages currently used or developed within HHS related to PCOR objectives.[45] Another project supported by the OS-PCORTF is a pilot to explore a novel privacy-preserving method called split learning, which uses data from health information exchanges (HIEs) for COVID-19-focused PCOR without compromising patient privacy.[46] Researchers may refer to the OS-PCORTF for example methods and approaches.

Consideration 2. Consider the tradeoffs involved in taking action to improve one or more TAI principles within a project and document decisions regarding tradeoffs throughout the research lifecycle. (Fair/Impartial; Transparent/Explainable; Responsible/Accountable; Robust/Reliable; Privacy; Safe/Secure)

Research has highlighted that tradeoffs are involved when applying TAI, since taking action to improve one principle can make it more challenging to adhere to another. For example, removing key identifiers from a data set to improve privacy protections for individuals may result in removing variables relevant to subpopulations such as members of racial/ethnic minority groups, affecting both fairness and robustness.[47] Several informants noted that making an assessment of the priority TAI principles is a critical step in the research planning phase and that prioritization of principles is likely to shift depending on how the AI results are to be used.[27] Assessment and documentation of tradeoffs is important to identify lessons and guidance for future OS-PCORTF projects.

Consideration 3. Address and re-assess AI solutions for trustworthiness in every phase of the research lifecycle. (Fair/Impartial; Transparent/Explainable; Responsible/Accountable; Robust/Reliable; Privacy; Safe/Secure)

Consideration of the TAI principles must occur in all phases of AI-enabled research. Not only does failing to address AI principles in any phase risk overall negative consequences,[3] but failing to address principles in an earlier phase can make it challenging to adequately address them fully in a later phase. Informants also stressed the need to repeatedly reassess AI solutions for trustworthiness in each phase to account for unexpected shifts in risks.[3]

Phase 1: Plan

Consideration of TAI principles at the planning stage of a study will lay the foundation for continual reassessment throughout the research lifecycle.

Consideration 4. Determine whether AI is appropriate for the research questions you are trying to answer, before beginning the project. (Fair/Impartial; Responsible/Accountable)

Understanding and clearly describing the problem an AI solution seeks to address is the first step for developing and using TAI in research. A growing body of literature on problem formulation in data science addresses identifying the research problem to be solved and how to train an algorithmic model to achieve those aims.[48] As detailed in Consideration 5, a peer review process or consultation with a steering committee or other governing body can support careful review of ethical issues.

A groundbreaking study that illustrates this consideration found evidence of racial bias in a widely used health care algorithm that assessed health needs by using health expenditures as a proxy for need.[49] The study found that because less money was spent on Black patients who had the same level of medical complexity and need as White patients, the algorithm had high predictive value of its target variable of

health cost. However, what this approach ignored was that Black patients have lower health expenditures than White patients, despite having higher morbidity and mortality rates. As a result, this cost-based algorithm would perpetuate the health care utilization resource bias observed in the data when it was used to direct health care resources.

Consideration 5. Develop a Steering Committee or Technical Expert Panel when leveraging AI solutions. The Committee or Panel should comprise experts in AI, data management, and information technology, as well as representatives of the patients and/or communities impacted by the work. Where possible, consider the inclusion of experts that are also members of the affected communities. (Fair/Impartial; Transparent/Explainable; Responsible/Accountable; Robust/Reliable; Privacy; Safe/Secure)

Establishing a clear advisory body or governance structure at the start of the research process facilitates adherence to TAI principles in all phases of the research lifecycle. Steering committees or technical expert panels could advise on the creation of research questions, the identification and review of appropriate data sets, analytic methods and tools, and clear, detailed data protection plans.

For the advisory function, our environmental scan identified the need for close collaboration with cross-disciplinary stakeholders, including those with expertise in areas such as mitigating bias, statistical or ML techniques, making data "AI ready," risk mitigation, privacy, security, and health system strengthening.[50, 20, 38] However, several key informants noted that building a bench of experts in the predictive AI field, as in any emergent field, is an ongoing enterprise-wide challenge. Additional workforce development and training will be required to help ensure that AI is used effectively, responsibly, and appropriately in research.

> "When you're starting to plan the program or plan the work that you're going to do as a researcher, having representatives and a steering committee that are from the communities of interest is critical. They see things that we don't see."
>
> **-Federal Key Informant**

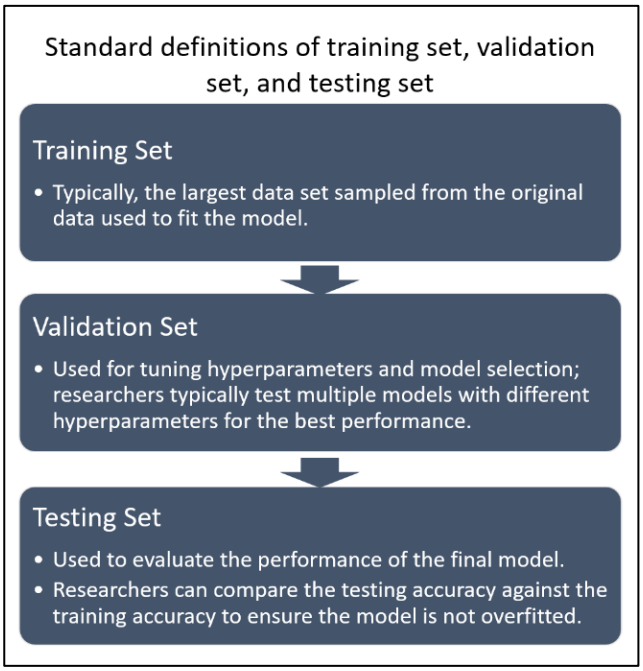
Researchers should consider the perspectives of the community, patients, health system leaders, and end-users such as clinicians affected by AI use, not only to identify potential biases in an AI tool but also to promote trust. Several key informants emphasized this point, particularly in the context of PCOR research. Patients can be engaged in steering committees, advisory boards, focus groups, or other governance structures. In a community governance model, for example, patients or research participants can be consulted on topics such as assessing the risks of data usage, identifying methods to minimize potential harms, and ensuring the priority population benefits from the results of the research.[42] The National Institutes of Health (NIH) Artificial Intelligence/Machine Learning Consortium to Advance Health Equity and Researcher Diversity (AIM-AHEAD) program provides a good model for increasing community engagement in research that leverages AI.[51]

Consideration 6. Enhance transparency and protect privacy by implementing clear, thorough data consent procedures that explicitly address the use of patient data in AI models. (Transparent/Explainable; Responsible/Accountable; Privacy)

Researchers involved in AI development, and who have access to patients whose data are being leveraged, should communicate with patients about how AI tools work in general and how their data may be used in training AI tools to make decisions. Patients have the right to a plain-language explanation about how their data will be used in an AI-enabled research project, potential harms that could result, and their right to withdraw their data.[52]

Informed consent procedures must use accessible language to make patients aware of key factors that may affect their willingness to have their data included in research that uses AI methods or tools. Other strategies to enhance the consent process can include engaging patients or participants through a variety of messaging formats such as short quizzes, games, or visualizations that lay out how AI tools work in general and how their data may be used in training other AI tools; such engagement may increase comprehension and trust in the research process.[53]

Because AI requires enormous volumes of data, it may not be feasible to gather informed consent from all patients. Further, individuals who are willing and able to provide informed consent may differ from individuals who will not or cannot provide informed consent, making the data less representative. Therefore, health care research organizations should identify when broad consent may be obtained (in place of informed consent) when data are to be used for secondary research. Broad consent includes most of the general requirements of informed consent but allows for broad categories of the types of research that may be conducted.[54]



Phase 2: Acquire

We identified three considerations for TAI that inform researchers' acquisition of data sets for use in training, validation, and testing of AI (see text box and Appendix E for definitions).[55]

Consideration 7. Determine the appropriate volume and quality of data to support the identified problem and AI application, when identifying data sets. (Fair/Impartial; Robust/Reliable)

Researchers should assess whether appropriate data are available to answer their research questions using AI methods. As several key informants noted, this critical step can be challenging due to limited metadata available on data sets to assess data quality and bias. Among the few data quality frameworks available to assess "fit-for-purpose," or data appropriateness for the intended purpose, the 3x3 Data Quality Assessment (DQA) Framework offers a set of guidelines to assess data quality for a given patient, variable, and time.[56] The harmonized DQA performs data checks on three categories—conformance, completeness, and plausibility[57]—and related subdomains.

Consideration 8. Assess whether the selected data sets represent the population being studied, when leveraging secondary data sets such as clinical registries, claims, or EHR data. (Fair/Impartial)

One cause of algorithmic bias is the use of biased data to train algorithms. Such bias can result in producing outputs that are systematically unfair to certain groups of people, for example, through age discrimination, racial bias, or gender bias.[58] Algorithmic fairness can be increased through the appropriate selection of representative data to train algorithms and by using statistical and ML techniques for algorithms that can mitigate bias.[59]

For example, researchers can conduct a data bias review by employing statistical techniques such as multivariate analysis.[60] They may consider bias review metrics such as false positive and negative rates to measure model fairness.[61] Several key informants noted that despite the increasing availability of methods and tools to assess bias, there is a need for: 1) guidance on the most appropriate metrics and measures to assess bias, and 2) consistent application of those metrics to support comparability.

> "Really, the important area for research on patients is enabling the need to partner with those communities to create data sets to make them available for model development and model tuning that traditionally are unreached populations: inner-city African Americans, rural White Appalachia, Rio Grande, Southwestern Hispanic populations...the list goes on."
>
> **-Non-Federal Key Informant**

Consideration 9. Explore the use of multiple, diverse, and high-quality data sources that support the identified use case for the AI model, including validation and training data sets. (Fair/Impartial; Transparent/Explainable; Responsible/Accountable; Robust/Reliable; Privacy)

As described above, the training, validating, and testing of AI models must use high-quality data suitable for the AI use case and population of interest. Using multiple and diverse data sources during AI development can ensure generalizability and minimize bias in algorithms. When an AI algorithm fits too closely against a single data set, there is a risk that the output cannot be generalized to other data sets (that is, data overfitting has occurred). Conversely, the opposite risk (underfitting) may occur where there is not enough complexity in the model for a robust match, which can produce biased results.[62] Overfitting and underfitting can be assessed and potentially mitigated with reliable external data validation.[62] In addition, researchers may minimize such risks by using multiple merged data sets to train and validate the algorithm. However, researchers must assess the semantic compatibility of the multiple data sets prior to merging, and this assessment can be a challenge.

Synthetic data—including clinical data, radiological images, or even survey responses—may be used to supplement the volume and diversity of data needed for an AI model. Use of synthetic data can also protect patient privacy. However, key informants cautioned that synthetic data may not represent all characteristics of real-world data that an ML model uses, so that the applicability and appropriateness of synthetic data should be carefully considered for every use case.[63] Generally, synthetic data is best used for initial training of an AI system, in combination with retraining and testing with real-world data. In addition, synthetic data may be used during testing to see how a model reacts to different populations or patient cases.[64] Examples of publicly available synthetic data sets include MDClone[65] (a free and secure platform for using synthetic health care data) and Synthea[66] (an open-source, synthetic patient generator that includes medical history of synthetic patients). If multiple data sources cannot be acquired, steps can be taken to augment the data (see discussion below, under Phase 3: Prepare).

Phase 3: Prepare

The third phase involves cleaning, organizing, and augmenting (if needed) the data to prepare for algorithmic training, development, and analysis.

Consideration 10. Consider techniques and methods that augment the data used, especially in situations where multiple, diverse data sets cannot be acquired or the amount of data needs to be artificially increased. (Fair/Impartial; Transparent/Explainable; Robust/Reliable; Privacy)

Researchers can use data augmentation techniques when processing data to artificially increase and/or balance the training data to make them more representative of the population of interest.[67] Several

informants emphasized the importance of data augmentation techniques, including blending data sets to increase size, representativeness, and diversity. When applied correctly, data augmentation can offer benefits including improved model accuracy, reduced data collection costs, prediction of rare events, and prevention of data privacy issues.

In addition, researchers can augment synthetic data to make the data representative of the population of interest while protecting patient privacy, with the understanding that augmentation still may not include all the characteristics needed to mimic real-world data.

In all cases, researchers must document data augmentation techniques clearly, to ensure transparency regarding the underlying data the algorithm uses.[68, 69]

Consideration 11. Consider having processes and protocols in place to reduce measurement error, missing data, and selection bias, any or all of which may occur during data collection. (Fair/Impartial; Robust/Reliable)

Researchers conducting primary data collection can mitigate measurement error that can result in biased data by taking multiple measurements of the same variable and collecting data with precision. Depending on the context of use, missing data can be addressed through imputation or modeling the missingness.

To address selection bias (that is, when the study sample does not accurately represent the target population) researchers can use sophisticated ML methods to develop the model and then sensitivity analyses can be conducted using simpler methods. For example, researchers can "up-sample" or "down-sample" the data according to weights. To support transparency, researchers should document and report whichever method they used for addressing missing data or selection bias. In reporting, researchers should always provide a justification for why the AI model used is appropriate for the sample size.[70]

Phase 4: Analyze

In the context of research that employs AI, the analyze phase includes analyzing and assessing outputs of an AI model.[71]

Consideration 12. Evaluate and test models for model performance, efficacy, accuracy, and adherence to principles using specific metrics, and continue to monitor performance over time. (Fair/Impartial; Transparent/Explainable; Responsible/Accountable; Robust/Reliable)

> "Especially because of our position in the government, being in a position of public trust and wanting to do right by the people, having a really solid evaluation plan for a model is super critical... if we don't have a really rigorous and carefully established design for evaluating the models, then [the rest] doesn't matter… what matters is that we're able to document how the model performs, not necessarily how it works."
>
> **-Federal Key Informant**

Having an evaluation plan in place is critical to ensuring the explainability and robustness of AI models and reveals opportunities for making refinements to reduce risk and mitigate bias. Evaluation of model performance (for example, efficacy and accuracy) involves assessing the AI model's ability to produce the correct output for a given input.[72]

To monitor model performance over time, and to support comparisons across models, researchers need to define clear, specific evaluation metrics that can be applied consistently. Key informants discussed

the importance of evaluation metrics and acknowledged the lack of guidance on metrics, consistent with our findings from the environmental scan. Groups like CHAI are working to develop assurance standards for trustworthy principles that may inform metric selection.[28] One key informant suggested that since there is no consensus yet on specific metrics, a starting point may be to develop metrics that assess models' performance and bias on a graded scale of low-medium-high.

Ground truth data sets can be used to train and test an AI algorithm against a labeled data set that reflects the reality that the research team wants to model.[73] The labeled data sets are developed by qualified experts. Researchers could use a ground truth data set in multiple algorithms within a field of study to test and refine metrics for model robustness and fairness and to compare outputs across studies.

All informants agreed that documenting and reporting model performance based on evaluation results is critical to promote transparency in AI use for research. Model cards are short documents that accompany AI models and can be used to convey details regarding a model's intended use case, how models were evaluated and evaluation results.[74] See Phase 6: Share Results/Reuse for additional tools and considerations for reporting results.

Consideration 13. Test and assess algorithms and their outcomes for risk of bias using appropriate analytic tools and techniques, when using AI. (Fair/Impartial)

Algorithmic fairness can be increased by employing statistical and ML techniques that can mitigate bias.[59] One commonly cited technique to assess algorithmic bias is to test for counterfactual fairness, or whether an algorithmic outcome or decision produced for an individual belonging to a particular demographic group would be the same for that individual in the real-world as in a counterfactual world where the individual belonged to a different demographic group.[19] Another technique is adversarial learning, which can mitigate bias from data sets by tricking a model with inputs that are considered stereotyped, to test the model's ability to predict the stereotype.[75]

Algorithmic models may underperform for subpopulations (for example, gender, race) due to limited available data. For this reason, it is important to identify subpopulations of interest and conduct analyses to detect disparities between less and more socially advantaged populations across model performance metrics, patient outcomes, and resources, to highlight and mitigate risk of bias.[76, 36] See text box for more examples of tools to assess fairness or risk of bias.[61, 77, 71]

Example Tools to Assess Fairness

- IBM has produced a Python toolkit for algorithmic fairness, "AI Fairness 360," that includes over 70 fairness metrics for data sets and models and provides 10 algorithms to mitigate bias in data sets and models.[61]
- FairMLHealth is a GitHub page of tools and tutorials for variation analysis in health care machine learning.[77]
- QUADAS-2 is designed to assess the source of bias in diagnostic accuracy studies through signaling risk questions within four domains: patient selection, index tests, reference standard, and flow and timing of patients through the study.[71]

Phase 5: Maintain/Update/Preserve

Phase 5 involves activities specific to AI-enabled research, including monitoring the performance of AI tools and maintaining data integrity.

Consideration 14. Monitor and maintain deployed systems continuously to identify and address risks and adverse events. (Privacy; Safe/Secure)

Health data must be properly managed to protect patients' privacy and security while maintaining accessibility for relevant stakeholders. Systems should be implemented to routinely ensure data are not breached, and researchers should account for varying levels of data sensitivity depending on context.[78] Further, once an AI system has been deployed, it must be monitored and maintained to identify and address risks and adverse events related to patient safety, data privacy, and security.

For patient safety, the researcher community should consider systematic approaches to monitoring through algorithmovigilance, or the science related to the "evaluation, monitoring, understanding, and prevention of adverse effects of algorithms in health care."[79] Algorithmovigilance entails careful assessment of algorithms during development and pre-deployment phases as well as systematic surveillance post-deployment. It considers how performance of algorithms may change as an algorithm is deployed with different data, in different settings, and at different times.

Security Requirements Engineering (SRE) is an area of software engineering that provides mechanisms to address the security, safety, risk, and vulnerability of health IT systems; SRE can help ensure AI safety and security.[80] Among the several SRE methods that can be applied to AI, the STORE methodology—a 10-step method that provides organizations with standard security practices and infrastructure—was found to be an effective approach for trustworthy health care software development.[80] A limited number of specific tools also exist to ensure safe and secure AI data sets. One such tool is IBM's "ART: Adversarial Robustness Toolbox,"[81] a Python library on GitHub that is dedicated to providing developers with tools to defend and evaluate machine learning models against adversarial attacks.

Phase 6: Share Results/Reuse

The FAIR principles (that all research objects should be Findable, Accessible, Interoperable, and Reusable) serve as a core set of guidelines for researchers intending to "enhance the reusability" of their data.[82] Activities to help achieve the FAIR principles include sharing data through scholarly publications or presentations, as well as disseminating data through data repositories or software such as GitHub.

Consideration 15. Promote transparency by comprehensively reporting on the functionality, strengths, and weaknesses of an AI tool. (Transparent/Explainable)

Lack of transparency can yield algorithmic systems that are hard to control, monitor, and correct and that will likely result in lack of trust among key stakeholders including the public.

Explainable AI (XAI) is a set of approaches and methods that aim to provide more meaningful and transparent explanations about how AI algorithms work and how one's data are used[84, 85] and serves as a useful framework to promote transparency. Researchers who develop AI models should be explicit about the intent, inputs, outputs, capabilities, and limitations of any AI product, including potential underlying biases in the AI's output. Use of reporting tools can facilitate transparency in communicating the strengths and limitations of an AI application.[86]

> **Explainable AI (XAI)** is a set of frameworks, tools, processes, and methods that allows users to understand and trust the results that ML algorithms create. XAI can be used to describe an AI model, its anticipated impact, and potential bias.[83] Currently there are XAI toolkits available from multiple companies and organizations.

A range of tools support the reporting process, and researchers should consider which reporting framework is best suited to their project. The framework of "contestable" AI is especially relevant for PCOR. The approach proposes that AI decision-making be explained clearly enough so all relevant stakeholders—including patients and providers—can contest the decision of the system.[87] In addition, a variety of reporting checklists, protocols, and guidelines can support AI transparency by ensuring that reporting contains the information crucial to evaluating the validity of AI health care tools; see Appendix C for a non-exhaustive table of reporting checklists.[88, 89, 90, 91] Model cards are an example of a documentation framework that promotes understanding and transparency of AI models in a way that is accessible to stakeholders with diverse backgrounds and varying needs to understand the model.[74] Several informants pointed out that transparent reporting on AI solutions facilitates open dialogue among researchers, including feedback loops between end-users and model developers that can support iterative improvements. Transparent reporting also avoids the "lemon market" of AI solutions, where potential users do not have enough information to assess the quality of the AI solutions they are considering for their research.[92]

6.3 Opportunities for the OS-PCORTF to Support Work that Promotes Adherence to the TAI Principles

We identified 14 potential opportunities for the OS-PCORTF to support improvements to tools, resources, and methods/techniques that facilitate adherence to HHS TAI principles. Each opportunity is tagged with the TAI principle(s) it addresses; see summary list in Exhibit 4.

Exhibit 4. Opportunities for the OS-PCORTF to Support Work Promoting Adherence to the TAI Principles **** The ordering of the opportunities does not reflect priority***
Governance Opportunity
• **Consider updating or amending products produced under the PCOR: Privacy and Security Blueprint, Legal Analysis and Ethics Framework for Data Use & Use of Technology for Privacy project to address the privacy, ethical, and legal considerations of including patient-level data in PCOR that integrate the use of AI algorithms or methods. (Privacy)**
Opportunities Related to Data
• **Support the development and adoption of high-quality, interoperable, standardized data sets. (Fair/Impartial; Robust/Reliable)**
• **Develop and pilot test methods for augmenting training data for use in AI to ensure that the data sets used are representative of the target population, minimize the risk of introducing bias in algorithms, and are regularly updated. (Fair/Impartial; Robust/Reliable)**
• **Engage in efforts that allow for further development and testing of synthetic data modules, such as Synthea, that can be used for training AI algorithms. (Fair/Impartial; Privacy)**
• **Conduct a proof-of-concept project to test a federated data model approach of training AI algorithms using multiple diverse training data sets from various sources. (Fair/Impartial; Robust/Reliable; Privacy; Safe/Secure)**
• **Support the leveraging of foundation models to develop models for research that could be adapted to many applications. (Robust/Reliable)**
Opportunities Related to Development of Tools and Resources
• **Develop resources and implementation guides to facilitate applications of privacy-enhancing technologies to enable secure and privacy-preserving methods of accessing and sharing data for use in AI. (Transparent/Explainable; Privacy; Safe/Secure)**

Exhibit 4. Opportunities for the OS-PCORTF to Support Work Promoting Adherence to the TAI Principles * *The ordering of the opportunities does not reflect priority*
• **Consider a meaningful aggregation of available tools to assess data sets for bias and provide guidance on which tools may be most appropriate for use in the PCOR context. (Fair/Impartial)**
• **Develop methodological guidance for the rigorous evaluation of AI/ML tools (Fair/Impartial; Transparent/Explainable)**
• **Consider supporting a comprehensive validation, review, and curation of existing AI reporting guidelines, checklists, and other frameworks to provide recommendations to researchers on the tools to leverage. (Responsible/Accountable; Transparent/Explainable)**
• **Develop resources that help PCOR researchers document and explain elements of AI models to non-technical users, including patients. (Responsible/Accountable; Transparent/Explainable)**
• **Create and maintain an inventory of all AI-related efforts undertaken in the OS-PCORTF portfolio. (Fair/Impartial; Transparent/Explainable; Responsible/Accountable)**
• **Develop an AI Research Core that PCOR researchers can reference for trusted tools and resources. (Robust/Reliable; Responsible/Accountable)**
• **Provide a forum for PCOR researchers to discuss what tools they have used to address TAI. (Responsible/Accountable; Transparent/Explainable)**

Governance Opportunity

Consider updating or amending products produced under the PCOR: Privacy and Security Blueprint and Legal Analysis and Ethics Framework for Data Use & Use of Technology for Privacy[93] projects, to address the privacy, ethical, and legal considerations of including patient-level data in PCOR that integrates the use of AI algorithms or methods. (Privacy)

Key informants shared that the rapid development of AI-enabled research requires updated, clear guidance on governance. One key informant specifically noted the need to examine the process by which patients consent to have their personal data included in research that leverages AI. An OS-PCORTF project, led by the Office of the National Coordinator for Health Information Technology (ONC) and the U.S. Centers for Disease Control and Prevention (CDC), created two products that could be revisited to determine if new or updated guidance could be provided in the context of AI use:

a. **Legal and Ethical Architecture for PCOR Data** describes a legal and ethical architecture to enable robust PCOR. It includes a collection of tools and resources that help researchers navigate legal requirements related to using data for PCOR.[94]
b. **Legal and Ethical Framework to Use CDC Data for Patient-Centered Outcomes Research** offers a legal and ethical framework to navigate challenges in allowing CDC's data to be used for PCOR.[95]

Opportunities Related to Data

Support the development and adoption of high-quality, interoperable, standardized data sets. (Fair/Impartial; Robust/Reliable)

Increased access to high-quality, standardized data sets for training, testing, and validating AI models is critical to creating AI that is fair, representative, and patient-centered. Improved data interoperability supports broader access to diverse data for research. Ensuring interoperability requires leveraging standards—such as HL7's Fast Healthcare Interoperability Resources® (FHIR)—that facilitate the exchange of data between health systems regardless of how data are stored and common data models that contain a uniform set of metadata that can be shared across applications.[96, 19]

Key informants recognized the importance of leveraging data sets that align with the FAIR principles of machine-ready data assets.[82] However, informants cautioned that selecting FAIR data exclusively for use in AI may eliminate useful and appropriate data sets for AI models and recommended taking action to make existing federal data sets consistent with FAIR principles by improving metadata and use of standards.

The OS-PCORTF has funded foundational work to support the creation of high-quality training data sets for ML models that predict mortality in the first 90 days of dialysis. This project produced a final report, an implementation guide, and publicly available code used to develop training data sets and ML models.[97] These efforts present opportunities for ASPE to expand on its work to enhance data sets for AI research and to make high-quality data available for PCOR.

Develop and pilot test methods to augment training data for use in AI to ensure the data sets used are representative of the target population, minimize the risk of introducing bias in algorithms, and are regularly updated. (Fair/Impartial; Robust/Reliable)

The training data for AI algorithms are critical to ensure outputs are fair and reliable, thus increasing trust in AI systems. To improve data quality, ASPE may explore new ways to supplement existing data sets to make them more representative and inclusive. Several methods and techniques exist to augment data to add variation and increase representativeness.[67] Generating more resources—to report on the effectiveness and usefulness of these techniques for researchers handling incomplete or insufficient data—will be important to improve algorithmic performance.

Engage in efforts that allow for further development and testing of synthetic data modules, such as Synthea, that can be used for training and testing AI algorithms.[98] (Fair/Impartial; Privacy)

Using synthetic data in research minimizes threats to data privacy and breaches of real patient data. Synthetic data may also be useful to improve representativeness of data sets used in AI models. There may be opportunities to expand Synthea synthetic data modules that can be used to evaluate algorithms. Researchers could benefit from a sandbox environment to test and iterate on AI algorithms and solutions. One informant also noted that other types of synthetic data resources could be developed and shared to support AI-enabled research, such as synthetic federal survey responses that could be used in NLP models.

Conduct a proof-of-concept project to test a federated data model approach to train AI algorithms using multiple diverse training data sets from different sources. (Fair/Impartial; Robust/Reliable; Privacy; Safe/Secure)

Research projects, including those funded by the OS-PCORTF, can maintain data privacy and security as they share AI models and train models collaboratively by not sharing the underlying data, and instead using decentralized data. For instance, in 2023, researchers at the University of Southern California proposed an architecture to address challenges in federated learning through principled data

integration and imputation techniques. This effort is still ongoing; it has the potential to make secure health data sharing less difficult across research teams and health systems.[99]

Support the leveraging of foundation models to develop models for research that could be adapted to many applications. (Robust/Reliable)

Foundation models are AI systems trained on large, unlabeled data sets (for example, structured and unstructured data from EHRs, imaging, laboratory results) that form the backbone of generative AI models.[100, 101, 102] Rather than build one model for a specific use case, which may be time- and labor-intensive, researchers may apply foundation models to different use cases, making the models an efficient resource to accelerate research. Careful assessment of the applicability and robustness of the foundation models for applies use cases is always warranted. A few key informants described the use of foundation models as important to democratize AI so that their applicability is broader and more flexible and this allows less reinforcement or additional model training to be done to achieve the goals of the application. The OS-PCORTF could consider funding projects that configure models for research leveraging foundation models, to expedite research to the analysis and evaluation phases.

Opportunities Related to Development of Tools and Resources

Develop resources and implementation guides to facilitate applications of privacy-enhancing technologies to enable secure and privacy-preserving methods of accessing and sharing data for use in AI. (Transparent/Explainable; Privacy; Safe/Secure)

For AI, PETs and data minimizing methods such as de-identification and aggregation for certain model outputs can support design for privacy-enhanced AI systems. The OS-PCORTF has already supported efforts in this area[45] and can continue to expand upon efforts to invest in novel privacy-enhancing methods. Researchers can also use approaches like federated learning to share AI models remotely and train models collaboratively without sharing the underlying data, maintaining data privacy and security.[103] Blockchain technology has been used to encrypt data and protect privacy; its unique architecture can complement deep learning by serving as a form of privacy-preserving technology.[104, 105]

Consider a meaningful aggregation of available tools to assess data sets for bias and provide guidance on which tools may be most appropriate for use in the PCOR context. (Fair/Impartial)

Many open-source and proprietary tools and toolkits are available to assess the fairness of data sets and algorithms, including IBM's AI Fairness 360,[61] FairMLHealth,[77] and QUADAS-2,[71] among others. But one key informant stated that many researchers are uncertain about the efficacy of these tools and their appropriateness for specific use cases. ASPE could fund an OS-PCORTF project to review such tools for their relevance and appropriateness for use in PCOR.

Develop methodological guidance for the rigorous evaluation of AI/ML tools. (Fair/Impartial; Transparent/Explainable)

Several key informants noted that researchers would benefit from detailed guidance on which methods should be used to evaluate the quality of AI tools. Informants described a fundamental need for specific metrics that assess adherence to TAI principles, since there is no consensus-driven standard agreement in the field to define such metrics. The OS-PCORTF could support this critical area by working to identify and to promote appropriate metrics.

> "Having a clear set of transparent metrics on how you're judging the principles in a technical manner is, I think, really important… [metrics] focused on these principles needs to be technical in nature, have technical specificity, and be transparent about that, so it's not another black box, and you're not sure how you're being evaluated, or the models are being evaluated."
>
> **-Non-Federal Key Informant**

Consider supporting a comprehensive validation, review, and curation of existing AI reporting guidelines, checklists, and other frameworks to provide recommendations to researchers on the tools to leverage. (Responsible/Accountable; Transparent/Explainable)

Numerous reporting guidelines and checklists have been created, including Transparent Reporting of Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD)[106] and Prediction model Risk Of Bias Assessment Tool (PROBAST)[106] (see Appendix C). However, little information is available to researchers on which guideline may be appropriate in the PCOR context. In addition, the OS-PCORTF could explore how to implement the curated list of checklists or frameworks with a patient focus in support of PCOR.

Develop resources that help PCOR researchers document and explain elements of AI models to non-technical users, including patients. (Responsible/Accountable; Transparent/Explainable)

There are tools, checklists, and reporting guidelines for AI in research to guide researchers on elements of the AI model that should be explained (for example, how the model was created and validated, checks for bias) to enhance trust in the system (see Appendix C); however, PCOR researchers could benefit from resources that provide guidance on how to explain the elements of AI models to non-technical users, including patients. As transparency tools for AI systems and related documentation evolve,[27] the OS-PCORTF could develop and test transparency tools in cooperation with PCOR researchers who employ AI and non-technical users, to develop resources that can support PCOR researchers in explaining AI models to non-technical users.

Create and maintain an inventory of all AI-related efforts undertaken in the OS-PCORTF portfolio. (Fair/Impartial; Transparent/Explainable; Responsible/Accountable)

As part of Executive Order 13960, government agencies are strongly encouraged to create a publicly available agency inventory of AI use cases, which HHS has provided.[4] To support this effort, ASPE can maintain their own inventory of new and existing AI systems and the descriptive characteristics of the systems (for example, projects/offices creating or deploying the system, data sources, intended use and users, AI capabilities, and links to published reports).[107] Agency-specific inventories could serve as models for a more broad-based registry of AI solutions; one key informant described the precedent that inventory efforts could set for developing a national, government-run registry for AI models, similar to the Clinical Trials registry. PCOR researchers could search such a registry to identify models that fit their purpose and, ultimately, use AI models that are more transparent, fair, and responsible.

Develop an AI Research Core that PCOR researchers can reference for trusted tools and resources. (Robust/Reliable; Responsible/Accountable)

Research Cores are centralized, shared resources that provide access to vetted tools, services, and specialized expertise for scientific and clinical investigators in a shared field. They offer a starting point for researchers who may not have experience in a specialized area to find reputable solutions or to speak with experts who can provide services to advance their research.[108] For instance, the National Cancer Institute has created a Core for cancer-related topics such as clinical research support, flow cytometry, and genetics and genomics for investigators to use for cancer research.[109] Key informants suggested creation of Research Cores as an opportunity for OS-PCORTF to expand access to AI applications so that PCOR researchers without related formal training or education can still use AI responsibly. Each resource, tool, and service in the Core would be validated by AI experts as trustworthy and/or high-quality.

Provide a forum for PCOR researchers to discuss what tools they have used to address TAI. (Responsible/Accountable; Transparent/Explainable)

There is widespread availability of AI tools and a growing body of research on their adoption, yet the benefits and consequences of AI tools are not well understood.[110] Key informants suggested that OS-PCORTF could provide a forum for PCOR researchers to share information regarding the tools they have used so far, the pros and considerations of such tools, and opportunities to collaborate. Such a forum would allow researchers to assess tools that are vetted and to consider strengths and limitations before applying to their research. Information gathered from the forum could identify gaps and opportunities for the OS-PCORTF to support improvements to the tools, resources, and methods/techniques for adherence to HHS TAI principles.

Conclusion

As AI becomes more prominent in health care, the use of tools and methods that promote TAI in health care research is an increasing priority. This report highlights the value of implementing the six TAI principles outlined in the HHS TAI Playbook when using AI for health care research that includes PCOR; the principles call for AI to be fair/impartial, transparent/explainable, responsible/accountable, robust/reliable, privacy, and safe/secure. Our findings reveal that several implementation considerations still need to be investigated and clarified. Key informants singled out the fair/impartial principle as the most difficult principle to implement and conceptualize and the privacy principle as the most intuitive, with longstanding tools and resources to support implementation. All six TAI principles will require constant monitoring and evaluation throughout the research lifecycle, as well as input from a range of community and expert stakeholders, to maintain trust in the decisions made from AI outputs.

The considerations outlined in the report reflect the overarching feedback from key informants so that researchers would be better equipped to implement TAI principles with the development of specific, actionable guidance on a range of topics.

This report serves as a resource for applying the six TAI principles within the context of OS-PCORTF projects and PCOR research to inform policymaking.

Appendix A. Additional Detail on Methods

Exhibit A1. Environmental Scan PubMed Search Terms

Search Term Category	Example Terms*
Artificial Intelligence	Artificial intelligence [MeSH Major Topic]
TAI Principles	Trust, trustworthy, bioethics, ethical, ethics, principles, responsible **fair**, impartial, unbiased, nondiscrimination, equity **transparent**, explainable, disclosure, understandable, open-source **responsible**, accountable, governance, monitored **safe**, secure, risk management, resilient **privacy**, confidential, protections, sensitive, consent **robust**, reliable, accurate, effective, quality, consistent
Implement	Implementation, application, translation, intervention, algorithm
Evaluation	Assess, evaluate, review, audit, measure, checklist, tools, metrics
Patient-Centered Outcomes Research	Patient-centered care [MeSH Major Topic]

*Bolded terms are the six principles in the HHS TAI Playbook, with related terms on the same line.

Exhibit A2. Environmental Scan Inclusion and Exclusion Criteria

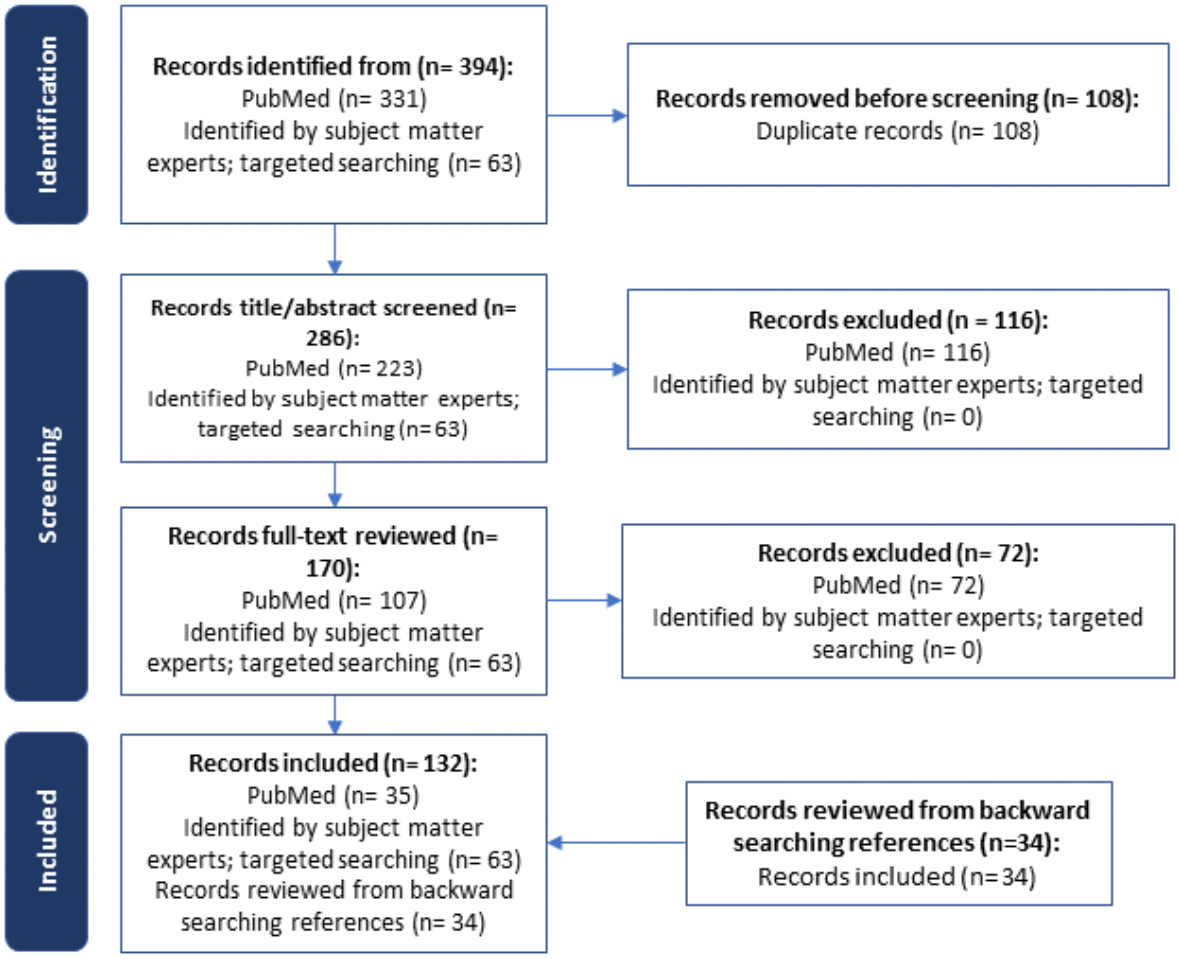
Category	Inclusion Criteria	Exclusion Criteria
Publication Year	2018 - present (last 5 years)	Prior to 2018
Document Type	**Gray literature**: Reports, evaluations, white papers, conference proceedings, case studies, fact sheets, issue briefs, blog if from a reputable expert **Peer review**: Theoretical articles, primary and secondary data analyses, scoping review, meta-analyses/systematic reviews	**Gray literature**: Opinion pieces
Sources	Academic, expert, evaluator	News outlet
Focus	Discusses consideration of or strategies for reviewing TAI principles in the context of PCOR or health care research	No discussion of reviewing TAI principles General discussion of TAI in relation to other sectors

Exhibit A3. Peer-Reviewed Literature Searches Conducted on PubMed

Search	Targeted Principle/ Focus	Search String	Filters Applied
1	Fair	("artificial intelligence"[MeSH Major Topic] AND ("trust"[Title/Abstract] OR "trustworthy"[Title/Abstract] OR "bioethics"[Title/Abstract] OR "responsible"[Title/Abstract] OR "ethical"[Title/Abstract] OR "ethics"[Title/Abstract] OR "governance"[Title/Abstract] OR "principles" [Title/Abstract]) AND ("implementation" OR "algorithm" OR "application" OR "translation" OR "intervention") AND ("assess" OR "evaluate" OR "review" OR "audit" OR "measure" OR "checklist") AND ("fair" OR "impartial" OR "unbiased" OR "nondiscrimination" OR "equity"))	2018-2023 English
2	Transparent	("artificial intelligence"[MeSH Major Topic] AND ("trust"[Title/Abstract] OR "trustworthy"[Title/Abstract] OR "bioethics"[Title/Abstract] OR "responsible"[Title/Abstract] OR "ethical"[Title/Abstract] OR "ethics"[Title/Abstract] OR "governance"[Title/Abstract]) AND ("implementation" OR "algorithm" OR "application" OR "translation" OR "intervention") AND ("assess" OR "evaluate" OR "review" OR "audit" OR "measure" OR "checklist") AND ("transparent" OR "explainable" OR "disclosure" OR "understandable" OR "open source"))	2018-2023 English
3	Responsible	("artificial intelligence"[MeSH Major Topic] AND ("trust"[Title/Abstract] OR "trustworthy"[Title/Abstract] OR "bioethics"[Title/Abstract] OR "responsible"[Title/Abstract] OR "ethical"[Title/Abstract] OR "ethics"[Title/Abstract] OR "governance"[Title/Abstract]) AND ("implementation" OR "algorithm" OR "application" OR "translation" OR "intervention") AND ("assess" OR "evaluate" OR "review" OR "audit" OR "measure" OR "checklist") AND ("responsible" OR "accountable" OR "traceable" OR "monitored" OR "governance"))	2018-2023 English
4	Safe	("artificial intelligence"[MeSH Major Topic] AND ("trust"[Title/Abstract] OR "trustworthy"[Title/Abstract] OR "bioethics"[Title/Abstract] OR "responsible"[Title/Abstract] OR "ethical"[Title/Abstract] OR "ethics"[Title/Abstract] OR "governance"[Title/Abstract]) AND ("implementation" OR "algorithm" OR "application" OR "translation" OR "intervention") AND ("assess" OR "evaluate" OR "review" OR "audit" OR "measure" OR "checklist") AND ("safe" OR "secure" OR "risk management" OR "resilient"))	2018-2023 English
5	Privacy	("artificial intelligence"[MeSH Major Topic] AND ("trust"[Title/Abstract] OR "trustworthy"[Title/Abstract] OR "bioethics"[Title/Abstract] OR "responsible"[Title/Abstract] OR "ethical"[Title/Abstract] OR "ethics"[Title/Abstract] OR "governance"[Title/Abstract] OR "principles"[Title/Abstract]) AND ("implementation" OR "algorithm" OR "application" OR "translation" OR "intervention") AND ("assess" OR "evaluate" OR "review" OR "audit" OR "measure" OR "checklist") AND ("privacy" OR "protections" OR "sensitive" OR "confidential" OR "consent"))	2018-2023 English

Search	Targeted Principle/ Focus	Search String	Filters Applied
6	Robust	("artificial intelligence"[MeSH Major Topic] AND ("trust"[Title/Abstract] OR "trustworthy"[Title/Abstract] OR "bioethics"[Title/Abstract] OR "responsible"[Title/Abstract] OR "ethical"[Title/Abstract] OR "ethics"[Title/Abstract] OR "governance"[Title/Abstract]) AND ("implementation" OR "algorithm" OR "application" OR "translation" OR "intervention") AND ("assess" OR "evaluate" OR "review" OR "audit" OR "measure" OR "checklist") AND ("robust" OR "reliable" OR "purposeful" OR "performance-driven" OR "accurate" OR "effective" OR "quality" OR "consistent"))	2018-2023 English
7	PCOR & AI	(patient centered care[MeSH Major Topic]) AND (artificial intelligence[MeSH Major Topic])	2018-2023 English

Exhibit A4. Article Selection Process



Appendix B. Key Informant Discussion Protocol

Introduction

1. I want to start by having you do a brief introduction of yourself. Can you please introduce yourself and briefly tell us about your work related to using AI in research?
2. Did you have the chance to review the information sheet? Do you have questions?

   As a brief overview, in September 2021 HHS published a Trustworthy AI (TAI) Playbook to provide guidance for HHS agencies on how to manage AI at all stages of the technology's lifecycle. The six TAI principles are fair/impartial, transparent/explainable, responsible/accountable, safe/secure, privacy, and robust/reliable.

3. [*For informants involved in research*] In what ways have TAI principles impacted how you conduct research?

Domain 1: Reactions to the List of Opportunities for OS-PCORTF to Support Alignment to TAI Principles

We'd like to first ask for your thoughts on the list of opportunities for OS-PCORTF in the information sheet that we shared with you before the call. These are potential ways to support PCOR researchers in aligning to HHS TAI principles. As a reminder, patient-centered outcomes research aims to generate high-quality evidence about the effectiveness of treatments, services, and other health care interventions on the full range of outcomes that patients, caregivers, clinicians, policymakers, and other stakeholders have identified as important.

4. Did any of the opportunities stand out to you as especially important? Why?
5. Of the opportunities described, which do you think would be important to focus on in the near-term (1-3 years) for improving alignment to TAI principles for patient-centered outcomes research?
6. Of the opportunities described, which do you think would be important to focus on in the long-term for improving alignment to TAI principles for patient-centered outcomes research?

   a. Why do you think this a long-term opportunity? Are there steps that can be taken in the short-term to help us achieve this opportunity?

7. After reading this list, are there other opportunities for activities or projects that OS-PCORTF could support related to implementing TAI in patient-centered outcomes research that you would add to the list?

   a. What TAI principle(s) would this opportunity address? How would this support use of TAI in patient-centered outcomes research?

Domain 2: Reactions to the List of Considerations for Project Teams Carrying Out OS-PCORTF Projects or PCOR Research

I would now like to turn to ask you about the list of considerations we provided for OS-PCORTF projects or PCOR researchers to align to trustworthy principles when using AI in their work. As a reminder, OS-PCORT projects focus on building data capacity for conducting patient-centered outcomes research (PCOR). One example of a project utilizing AI solutions to build data capacity is using machine learning to enable health information exchange for PCOR focused on COVID-19. The OS-PCORTF has also funded the creation of high-quality training data sets for machine learning for two use cases: kidney disease and drug resistance in patients infected with tuberculosis.

8. Are there any particular tools or resources that have been helpful in your work to align to the TAI principles?

    a. Are there any specific to evaluating or assessing PCOR or research projects for TAI principles?

9. In your AI-related work, which of the six principles have you found to be most challenging to adhere to and why?

    a. What types of resources would make it easier for you to adhere to these principles?

10. Are there any particular tools or resources for applying TAI principles that you have tried to use but faced barriers in their application? If so, please describe.
11. Are there considerations that you think are particularly relevant for the OS-PCORTF community (e.g., HHS partners) or PCOR researchers? Why?
12. Based on your experience, are there other considerations that we did not include that could be implemented by OS-PCORTF projects or PCOR researchers?

Conclusion

13. What future work do you think the OS-PCORTF can support in the area of TAI?
14. Is there anything else you would like to share about implementing or assessing trustworthy principles in PCOR?

Those are all the questions that we had for you today. Thank you again for your time and insights. Please do not hesitate to reach out if you have further questions about this project.

Appendix C. Reporting Checklists

Exhibit C1. Available Reporting Checklists and Protocols

Name	Description	Use
Transparent Reporting of Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD)[106]	TRIPOD is a 22-item checklist that includes items essential to the transparent and accurate reporting of **studies that develop or validate multivariable clinical prediction models**. TRIPOD-AI is an extension of the TRIPOD statement that was developed specifically for use with prediction model studies that leverage AI and machine learning techniques.	TRIPOD-AI provides researchers leveraging AI with guidelines to help them report details that will be needed for other researchers to evaluate their study's quality and interpret findings.
Prediction model Risk of Bias Assessment Tool (PROBAST)[106]	PROBAST works **to assess the risk of bias in and improve reporting of machine learning-based on multivariable prediction model studies for diagnosis and prognosis**. It uses 20 questions organized across four domains (participants, predictors, outcomes, and analysis) as the basis of its standardized tool for bias evaluation.	It serves as a tool for researchers to appraise, conduct, and analyze machine learning-based prediction model studies.
Standards for Reporting of Diagnostic Accuracy Standards (STARD)-AI[111]	First formulated in 2000, this protocol was designed to standardize **comparative studies of new or alternative diagnostic tests against an established reference standard** and was updated in 2021 to address key considerations for AI interventions. Some of the topics covered in the checklist include data processing methods, AI index test development methods, fairness metrics, explainability, and human-AI index tests. The goal is to generate a list of minimal essential items that should be reported in all AI diagnostic test accuracy studies.	STARD-AI supports researchers in appraising the quality and comparing the diagnostic test accuracy of AI models reported in scientific studies. **Note:** If the study is focusing on multivariable prediction models (for example, time to event predictions), the TRIPOD-AI may be more appropriate.
Consolidated Standards of Reporting Trials-Artificial Intelligence (CONSORT-AI)[112] **Extension**	Among the earliest of such protocols, introduced in 1996 (with further updates in 2001 and 2010), CONSORT aimed to specify reporting guidelines for parallel group randomized controlled trials (RCTs). The CONSORT-AI extension is a new reporting guideline **for clinical trial reports of interventions with an AI component**. It includes 14 new items with an AI focus, to be routinely reported in addition to the CONSORT 2010 items.	It supports researchers, editors, and peer reviewers in understanding, interpreting, and critically appraising the quality of clinical trial design and risk of bias in the reported outcomes.
Consolidated Standards of Reporting Trials of Electronic and mobile Health Applications and Online Telehealth (CONSORT-EHEALTH)[113]	The CONSORT-EHEALTH extension aims to capture the unique challenges of **reporting eHealth and mHealth RCTs**, particularly related to details that support reproducibility, theory-building, and implementation in other settings. The checklist includes 17 items that are considered "essential" and 35 subitems considered "highly recommended." Authors must address each of the items on the checklist.	This is useful for researchers of eHealth and mHealth interventions as well as researchers who use web-based recruitment or data collection methods. Many elements can be applied to evaluation reports, not just RCTs.

Name	Description	Use
Transparent Reporting of Evaluations with Nonrandomized Designs (TREND) Statement[114]	This is a 22-item checklist to guide **transparency in the reporting of non-RCTs**. The checklist complements CONSORT (described above). Designed for behavioral and public health intervention evaluations, the checklist focuses on the description of the intervention, description of the comparison condition, reporting of outcomes, and design information to assess possible biases.	Researchers, funding agencies, journal editors, and reviewers can use the statement as a guide when designing evaluation studies, reporting evaluation results, and reviewing manuscripts.
Standard Protocol Items: Recommendations for Intervention Trials-Artificial Intelligence (SPIRIT)-AI Extension[115]	SPIRIT was developed in 2013 to improve the completeness of clinical trial protocol reporting. Recognizing that AI interventions must undergo rigorous evaluations, researchers developed the SPIRIT-AI extension in tandem with CONSORT-AI to serve as a **reporting guideline for clinical trial protocols evaluating interventions with an AI component**.	SPIRIT-AI supports editors and peer reviewers to understand, interpret, and critically appraise the design and risk of bias for a planned clinical trial.

Appendix D. Table of Acronyms

Acronym	Description
AI	Artificial intelligence
AIM-AHEAD	Artificial Intelligence/Machine Learning Consortium to Advance Health Equity and Researcher Diversity
ASPE	Assistant Secretary for Planning and Evaluation
CDC	Centers for Disease Control and Prevention
CHAI	Coalition for Health Artificial Intelligence
CONSORT-AI	Consolidated Standards of Reporting Trials-Artificial Intelligence
DQA	Data Quality Assessment
EHR	Electronic health record
FAIR	Findable, Accessible, Interoperable, and Reusable
HHS	Health and Human Services
HIE	Health information exchange
ML	Machine learning
NAM	National Academy of Medicine
NIH	National Institutes of Health
NIST	National Institute of Standards and Technology
NLM	National Library of Medicine
NLP	Natural language processing
NORC	NORC at the University of Chicago
OCAIO	Office of the Chief AI Officer
OS-PCORTF	Office of the Assistant Secretary Patient-Centered Outcomes Research Trust Fund
PCOR	Patient-Centered Outcomes Research
PET	Privacy-enhancing technologies
RCT	Randomized controlled trials
SRE	Security Requirements Engineering
TAI	Trustworthy artificial intelligence
TREND	Transparent Reporting of Evaluations with Nonrandomized Designs
TRIPOD	Transparent Reporting of Multivariable Prediction Model for Individual Prognosis or Diagnosis
XAI	Explainable artificial intelligence

Appendix E. Glossary of Terms

Artificial Intelligence: The capability of computer systems to "perform tasks normally requiring human intelligence."[8]

Algorithm: A process or set of instructions that "will help calculate an answer to a problem," especially when given to a computer.[116]

Algorithmovigilance: The science related to the "evaluation, monitoring, understanding, and prevention of adverse effects of algorithms in health care."[79]

Bias: An error that can occur in an artificial intelligence model if the model's results are systematically prejudiced by its training data.[117]

"Black Box" Algorithms: Algorithms that "humans cannot survey," since they typically "do not follow well-defined rules" and are comprised of "opaque systems that no human or group of humans can closely examine."[118]

Contestable Artificial Intelligence: Artificial intelligence systems that are "open and responsive to human intervention throughout their lifecycle, not only after an automated decision has been made, but also during its design and development."[119]

Data Augmentation: A series of techniques that address the problem of limited data by "artificially increasing the amount of data by generating new data points from existing data," which can involve making changes to existing data or leveraging deep learning models to produce new data models.[120]

Data Set: A collection of separate but related sets of information that can be manipulated as a single unit by a computer.[121]

Deep Learning: A type of machine learning that leverages multilayered neural networks to simulate how the human brain behaves. These layers allow an algorithm to "learn from large amounts of data" and strengthen the algorithm's accuracy.[122]

Explainable AI (XAI): A set of frameworks, tools, processes, and methods that allow users to understand and trust the results created by machine learning algorithms. XAI can be used to describe an AI model, its anticipated impact, and potential bias.[83]

Foundation Model: A model that is trained on broad, unlabeled data that can be adapted to a wide range of tasks. Types of foundation models include generative AI (see definition below) and large language models.[100]

Generative AI: Deep learning models that are able to generate text, images, video, and other content by "identifying patterns in large quantities of training data, and then creating original material that has similar characteristics."[117] The popular AI model ChatGPT is an example of a generative AI tool.

Hyperparameter: The parameters whose values control the learning process of the algorithm. Hyperparameters are set before training a model so that the model cannot change its values during learning/training.[123]

Machine Learning: A sub-field of artificial intelligence that involves "the use and development of computer systems that are able to learn and adapt without following explicit instructions by using algorithms and statistical models to analyze and draw inferences from patterns in data."[9]

Metadata: Information describing the "characteristics of data."[124] It can include information about the content and context of a data set and information used to manage data.[125]

Natural Language Processing: A sub-field of artificial intelligence that works to "enable computers to process human language in the form of text or voice data and to 'understand' its full meaning."[126]

Overfitting: A modeling error that occurs when a statistical model fits exactly against a minimal set of training data. This makes the model "unable to perform accurately against unseen data, defeating its purpose."[127]

Privacy-Enhancing Technologies: Digital tools that "allow information to be collected, processed, analyzed, and shared while protecting data confidentiality and privacy."[128]

Representative Data: Data that includes accurate information on the communities that it may impact, which is critical for ensuring the effectiveness of AI algorithms.[129]

Synthetic Data: Computer-generated information that is used to "augment or replace real data to test and train artificial intelligence models."[130]

Testing Data: Data that is used after a machine learning model is built to "evaluate the performance and progress of [the] algorithms' training and adjust or optimize it for improved results."[131]

Training Data: An initial, large data set that is used to teach a machine learning model to "recognize patterns or perform your criteria."[131]

Trustworthy AI: The "design, development, acquisition, and use of AI in a manner that fosters public trust and confidence while protecting privacy, civil rights, civil liberties, and American values, consistent with applicable laws."[3]

Underfitting: An error in that occurs when a data model is overly simple or requires additional training time, which causes the model to be "unable to capture the relationship between the input and output variables accurately, generating a high error rate on both the training data set and unseen data."[132]

References

[1] Prem E. From ethical AI frameworks to tools: a review of approaches. *AI and Ethics*. 2023;3(3):699-716. doi:10.1007/s43681-023-00258-9

[2] Zhou J, Chen F, Berry A, Reed M, Zhang S, Savage S. A survey on ethical principles of AI and implementations. Paper presented at: 2020 IEEE Symposium Series on Computational Intelligence; December 2020; Canberra, Australia. doi: 10.1109/SSCI47803.2020.9308437

[3] U.S. Department of Health and Human Services. Trustworthy AI (TAI) playbook. HHS. September 2021. Accessed July 12, 2023. Retrieved from: https://www.hhs.gov/sites/default/files/hhs-trustworthy-ai-playbook.pdf

[4] Executive Office of the President. Executive order 13960, Promoting the use of trustworthy artificial intelligence in the federal government. The Federal Register. December 3, 2020. Accessed July 11, 2023. Retrieved from: https://www.federalregister.gov/documents/2020/12/08/2020-27065/promoting-the-use-of-trustworthy-artificial-intelligence-in-the-federal-government

[5] Kargl M, Plass M, Müller H. A literature review on ethics for AI in biomedical research and biobanking. *Yearb Med Inform*. 2022;31(1):152-160. doi:10.1055/s-0042-1742516

[6] Patient-Centered Outcomes Research Institute. Patient-centered outcomes research. PCORI. Accessed July 12, 2023. Retrieved from: https://www.pcori.org/research/about-our-research/patient-centered-outcomes-research

[7] Office of the Assistant Secretary for Planning and Evaluation. Office of the Secretary Patient-Centered Outcomes Research Trust Fund strategic plan, 2020-2029. ASPE. Accessed July 12, 2023. Retrieved from: https://aspe.hhs.gov/sites/default/files/documents/50ed989d0b92e5d9f20952bf357bd0ac/os-pcortf-2020-2029-strategic-plan-infographic.pdf

[8] Office of the Chief Information Officer, U.S. Department of Health and Human Services. Artificial intelligence (AI) at HHS. HHS. July 29, 2022. Accessed July 13, 2023. Retrieved from: https://www.hhs.gov/about/agencies/asa/ocio/ai/index.html

[9] AI and Machine Learning Platform Integration. IBM. Accessed July 13, 2023. Retrieved from: https://www.ibm.com/docs/en/cloud-paks/1.0?topic=cloudpaks_start/ibm-process-mining/user-manuals/ai_ml_platformintegration/introduction.htm

[10] Predictive Analytics. IBM. Accessed July 12, 2023. Retrieved from: https://www.ibm.com/analytics/predictive-analytics

[11] Office of the Assistant Secretary for Planning and Evaluation. Building data capacity for patient-centered outcomes research: Office of the Secretary Patient-Centered Outcomes Research Trust Fund strategic plan, 2020-2029. September 2022. ASPE. Accessed August 31, 2023. Retrieved from: https://aspe.hhs.gov/sites/default/files/documents/b363671a6256c6b7f26dec4990c2506a/aspe-os-pcortf-2020-2029-strategic-plan.pdf

[12] Office of the Assistant Secretary for Planning and Evaluation. Utilizing natural language processing and machine learning to enhance the identification of stimulant and opioid-involved health outcomes in the National Hospital Care Survey. ASPE. Accessed July 13, 2023. Retrieved from: https://aspe.hhs.gov/identifying-stimulant-use-nhcs-using-nlp

[13] Office of the Assistant Secretary for Planning and Evaluation. Using machine learning techniques to enable health information exchange to support COVID-19-focused patient-centered outcomes research (PCOR). ASPE. Accessed July 13, 2023. Retrieved from: https://aspe.hhs.gov/using-machine-learning-techniques-enable-health-information-exchange-support-covid-19-focused

[14] Office of the Assistant Secretary for Planning and Evaluation. Training data for machine learning to enhance patient-centered outcomes research (PCOR) data infrastructure. ASPE. Accessed July 13, 2023. Retrieved from: https://aspe.hhs.gov/training-data-machine-learning-enhance-patient-centered-outcomes-research-pcor-data-infrastructure

[15] Toner, H. What are generative AI, large language models, and foundation models? Georgetown University Center for Security and Emerging Technology. May 12, 2023. Accessed July 12, 2023. Retrieved from: https://cset.georgetown.edu/article/what-are-generative-ai-large-language-models-and-foundation-models/

[16] Tang X. The role of artificial intelligence in medical imaging research. *BJR Open*. 2019;2(1):20190031. doi:10.1259/bjro.20190031

[17] Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. *Future Healthc J*. 2019;6(2):94-98. doi:10.7861/futurehosp.6-2-94

[18] Chowdhury M, Cervantes EG, Chan WY, Seitz DP. Use of machine learning and artificial intelligence methods in geriatric mental health research involving electronic health record or administrative claims data: a systematic review. *Front Psychiatry*. 2021;12:738466. doi:10.3389/fpsyt.2021.738466

[19] Weissler EH, Naumann T, Andersson T, et al. The role of machine learning in clinical research: transforming the future of evidence generation. *Trials*. 2021;22(1):537. doi:10.1186/s13063-021-05489-x

[20] Matheny M, Thadaney Israni S, Ahmed M, Whicher D. Artificial intelligence in health care: the hope, the hype, the promise, the peril. National Academy of Medicine. 2022. Accessed August 31, 2023. Retrieved from: https://nam.edu/wp-content/uploads/2019/12/AI-in-Health-Care-PREPUB-FINAL.pdf

[21] Bohr A, Memarzadeh K. The rise of artificial intelligence in healthcare applications. *Artificial Intelligence in Healthcare*. 2020;25-60. doi:10.1016/B978-0-12-818438-7.00002-2

[22] Askin S, Burkhalter D, Calado G, El Dakrouni S. Artificial intelligence applied to clinical trials: opportunities and challenges. *Health Technol (Berl)*. 2023;13(2):203-213. doi:10.1007/s12553-023-00738-2

[23] Härkänen M, Haatainen K, Vehviläinen-Julkunen K, Miettinen M. Artificial intelligence for identifying the prevention of medication incidents causing serious or moderate harm: an analysis using incident reporters' views. *Int J Environ Res Public Health*. 2021;18(17):9206. doi:10.3390/ijerph18179206

[24] Soni H, Ivanova J, Wilczewski H, et al. Virtual conversational agents versus online forms: patient experience and preferences for health data collection. *Front Digit Health*. 2022;4:954069. doi:10.3389/fdgth.2022.954069

[25] Jain A, Brooks JR, Alford CC, et al. Awareness of racial and ethnic bias and potential solutions to address bias with use of health care algorithms. *JAMA Health Forum*. 2023;4(6):e231197. 2023 Jun 2. doi:10.1001/jamahealthforum.2023.1197

[26] Office of Science and Technology Policy, The White House. Blueprint for an AI Bill of Rights. The White House October 2022. Accessed July 24, 2023. Retrieved from: https://www.whitehouse.gov/ostp/ai-bill-of-rights/

[27] National Institute of Standards and Technology. Artificial intelligence risk management framework (AI RMF 1.0). NIST. January 2023. Accessed July 6, 2023. Retrieved from: https://www.nist.gov/itl/ai-risk-management-framework

[28] Coalition for Health AI. Blueprint for trustworthy AI implementation guidance and assurance for healthcare. CHAI. April 2023. Accessed July 12, 2023. Retrieved from: https://coalitionforhealthai.org/papers/blueprint-for-trustworthy-ai_V1.0.pdf

[29] National Academy of Medicine. Health care artificial intelligence code of conduct. NAM. Accessed July 13, 2023. Retrieved from: https://nam.edu/programs/value-science-driven-health-care/health-care-artificial-intelligence-code-of-conduct/

[30] Mittelstadt BD, Allo P, Taddeo M, Wachter S, Floridi L. The ethics of algorithms: mapping the debate. *Big Data & Society*. 2016;3(2). doi:10.1177/2053951716679679

[31] The National Library of Medicine. Research lifecycle. NLM. Accessed July 12, 2023. Retrieved from: https://www.nnlm.gov/guides/data-glossary/research-lifecycle

[32] Appari A, Johnson ME. Information security and privacy in healthcare: current state of research. *Int. J. Internet and Enterprise Management*. 2010;6(4): 279-314

[33] Shi S, He D, Li L, et al. Applications of blockchain in ensuring the security and privacy of electronic health record systems: a survey. *Computers & Security*. 2020;97. doi:10.1016/j.cose.2020.101966

[34] Kelly S. Rapid AI adoption could cause medical errors, patient harm, WHO warns, urging oversight. Healthcare Dive. Accessed July 13, 2023. Retrieved from: https://www.healthcaredive.com/news/WHO-artificial-intelligence-AI-caution/650538/

[35] Sunarti S, Fadzlul Rahman F, Naufal M, Risky M, Febriyanto K, Masnina R. Artificial intelligence in healthcare: opportunities and risk for future. *Gac Sanit*. 2021;35 Suppl 1:S67-S70. doi:10.1016/j.gaceta.2020.12.019

[36] Obermeyer Z, Nissan R, Stern M, Eaneff S, Bembeneck EJ, Mullainathan S. Algorithmic bias playbook. Chicago Booth Center for Applied Artificial Intelligence. June 2021. Accessed August 31, 2023. Retrieved from: https://www.chicagobooth.edu/research/center-for-applied-artificial-intelligence/research/algorithmic-bias

[37] Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. *Ann Intern Med*. 2018;169(12):866-872. doi:10.7326/M18-1990

[38] Siala H, Wang Y. SHIFTing artificial intelligence to be responsible in healthcare: A systematic review. *Soc Sci Med*. 2022;296:114782. doi:10.1016/j.socscimed.2022.114782

[39] Weinkauf D. Privacy tech-know blog: When worlds collide – the possibilities and limits of algorithmic fairness (part 1). Office of the Privacy Commissioner of Canada. April 5, 2023. Accessed July 31, 2023. Retrieved from: https://www.priv.gc.ca/en/blog/20230405_01/

[40] Al-Ruithe M, Benkhelifa E, Hameed K. A systematic literature review of data governance and cloud data governance. *Pers Ubiquit Comput*. 2019;23:839–859. https://doi.org/10.1007/s00779-017-1104-3

[41] Allen C, Des Jardins TR, Heider A, et al. Data governance and data sharing agreements for community-wide health information exchange: lessons from the beacon communities. *EGEMS (Wash DC)*. 2014;2(1):1057. doi:10.13063/2327-9214.1057

[42] Fisher S, Rosella LC. Priorities for successful use of artificial intelligence by public health organizations: a literature review. *BMC Public Health*. 2022;22(1):2146. doi:10.1186/s12889-022-14422-z

[43] Goldsteen A, Saadi O, Shmelkin R, Shachor S, Razinkov N. AI privacy toolkit. *SoftwareX*. 2023;22:101352. doi:10.1016/j.softx.2023.101352

[44] National Institute of Standards and Technology. Privacy framework. NIST. Accessed July 12, 2023. Retrieved from: https://www.nist.gov/privacy-framework

[45] Office of the Assistant Secretary for Planning and Evaluation. Evaluation of privacy-preserving record linkage solutions to broaden linkage capabilities in support of patient-centered outcomes research objectives. ASPE. Accessed July 27, 2023. Retrieved from: https://aspe.hhs.gov/evaluation-privacy-preserving-record-linkage-solutions-broaden-linkage-capabilities

[46] Office of the National Coordinator for Health Information Technology. Using machine learning techniques to enable health information exchange to support COVID-19-focused patient-centered outcomes research. ONC. Accessed July 27, 2023. Retrieved from: https://www.healthit.gov/topic/research-evaluation/using-machine-learning-techniques-enable-health-information-exchange

[47] Pujol D, Machanavajjhala A. Equity and privacy: more than just a tradeoff. *IEEE Security & Privacy*. 2021;*19*(6), 93-97. doi: 10.1109/MSEC.2021.3105773

[48] Passi S, Barocas S. Problem formulation and fairness. Paper presented at: The Conference on Fairness, Accountability, and Transparency; January 29-31, 2019; Atlanta, Georgia.

[49] Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 2019;366(6464):447-453. doi:10.1126/science.aax2342

[50] Reddy S, Allan S, Coghlan S, Cooper P. A governance model for the application of AI in health care. *J Am Med Inform Assoc*. 2019;27(3):491-497. doi:10.1093/jamia/ocz192

[51] Office of Data Science Strategy, National Institutes of Health. Artificial intelligence/machine learning consortium to advance health equity and researcher diversity (AIM-AHEAD). NIH. October 6, 2021. Accessed July 11, 2023. Retrieved from: https://datascience.nih.gov/artificial-intelligence/aim-ahead

[52] Andreotta AJ, Kirkham N, Rizzi M. AI, big data, and the future of consent. *AI Soc*. 2022;37(4):1715-1728. doi:10.1007/s00146-021-01262-5

[53] Jacobson NC, Bentley KH, Walton A, et al. Ethical dilemmas posed by mobile health and machine learning in psychiatry research. *Bull World Health Organ*. 2020;98(4):270-276. doi:10.2471/BLT.19.237107

[54] Maloy JW, Bass PF. Understanding broad consent. *Ochsner J.* 2020;20(1):81-86. doi:10.31486/toj.19.0088

[55] Kuhn M, Johnson K. *Applied Predictive Modeling*. 1st ed. Springer; 2013.

[56] Weiskopf NG, Bakken S, Hripcsak G, Weng C. A data quality assessment guideline for electronic health record data reuse. *EGEMS (Wash DC)*. 2017;5(1):14. doi:10.5334/egems.218

[57] Kahn MG, Callahan TJ, Barnard J, et al. A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. *EGEMS (Wash DC).* 2016;4(1):1244. doi:10.13063/2327-9214.1244

[58] Panch T, Mattie H, Celi LA. The "inconvenient truth" about AI in healthcare. *NPJ Digit Med.* 2019;2:77. https://doi.org/10.1038/s41746-019-0155-4

[59] Norori N, Hu Q, Aellen FM, Faraci FD, Tzovara A. Addressing bias in big data and AI for health care: a call for open science. *Patterns (NY).* 2021;2(10):100347. doi:10.1016/j.patter.2021.100347

[60] Defense Innovation Board, U.S. Department of Defense. AI principles: recommendations on the ethical use of artificial intelligence. DOD. October 31, 2019. Accessed on July 7, 2023. Retrieved from: https://media.defense.gov/2019/Oct/31/2002204458/-1/-1/0/DIB_AI_PRINCIPLES_PRIMARY_DOCUMENT.PDF

[61] AI Fairness 360. IBM. November 14, 2018. Accessed July 7, 2023. Retrieved from: https://www.ibm.com/opensource/open/projects/ai-fairness-360/

[62] He M, Li Z, Liu C, Shi D, Tan Z. Deployment of artificial intelligence in real-world practice: opportunity and challenge. *Asia Pac J Ophthalmol (Phila).* 2020;9(4):299-307. doi:10.1097/APO.0000000000000301

[63] Savage N. Synthetic data could be better than real data [published online ahead of print, 2023 Apr 27]. *Nature*. 2023;10.1038/d41586-023-01445-8. doi:10.1038/d41586-023-01445-8

[64] Chen RJ, Lu MY, Chen TY, et al. Synthetic data in machine learning for medicine and healthcare. *Nat Biomed Eng*. 2021;5:493–497. https://doi.org/10.1038/s41551-021-00751-8

[65] Overview. MDClone. Accessed July 10, 2023. Retrieved from: https://www.mdclone.com/about-mdclone

[66] Synthea. Synthetic patient generation. GitHub. Accessed July 10, 2023. Retrieved from: https://syntheticheath.github.io/synthea/

[67] Abels E, Pantanowitz L, Aeffner F, et al. Computational pathology definitions, best practices, and recommendations for regulatory guidance: a white paper from the Digital Pathology Association. *J Pathol*. 2019;249(3):286-294. doi:10.1002/path.5331

[68] Hernández-García A, König P. Further advantages of data augmentation on convolutional neural networks. In: Kůrková V, Manolopoulos Y, Hammer B, Iliadis L, Maglogiannis I., eds. *Lecture Notes in Computer Science*. Springer; 2018. doi:10.1007/978-3-030-01418-6_10

[69] Rommel C, Paillard J, Moreau T, Gramfort A. Data augmentation for learning predictive models on EEG: a systematic comparison. *J Neural Eng*. 2022;19(6). doi:10.1088/1741-2552/aca220

[70] Caliskan A. Detecting and mitigating bias in natural language processing. Brookings Institute. 2021. Accessed May 30, 2023. Retrieved from: https://www.brookings.edu/research/detecting-and-mitigating-bias-in-natural-language-processing/

[71] Whiting PF, Rutjes AW, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med*. 2011;155(8):529-536. doi:10.7326/0003-4819-155-8-201110180-00009

[72] Lebovitz S, Levina N, Lifshitz-Assaf H. Is AI ground truth really true? The dangers of training and evaluating AI tools based on experts' know-what. *MIS Quarterly*. 2021;45(3).

[73] Ground Truth. Domino Data Lab. Accessed July 11, 2023. Retrieved from: https://www.dominodatalab.com/data-science-dictionary/ground-truth

[74] Mitchell M, Wu S, Zaldivar A, et al. Model cards for model reporting. ArXiv. January 14, 2019. Accessed July 11, 2023. Retrieved from: https://arxiv.org/pdf/1810.03993.pdf

[75] Zhang BH, Lemoine B, Mitchell M. Mitigating unwanted biases with adversarial learning. In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18).* Association for Computing Machinery. 2018:335-340. doi:10.1145/3278721.3278779

[76] Rojas JC, Fahrenbach J, Makhni S, et al. Framework for integrating equity into machine learning models: a case study. *Chest.* 2022;161(6):1621-1627. doi:10.1016/j.chest.2022.02.001

[77] KenSciResearch. FairMLHealth. GitHub. Accessed July 11, 2023. Retrieved from: https://github.com/KenSciResearch/fairMLHealth/blob/integration/fairmlhealth/README.md

[78] Mani V, Kavitha C, Band SS, Mosavi A, Hollins P, Palanisamy S. A recommendation system based on AI for storing block data in the electronic health repository. *Front Public Health*. 2022;9:831404. doi:10.3389/fpubh.2021.831404

[79] Embi PJ. Algorithmovigilance—advancing methods to analyze and monitor artificial intelligence-driven health care for effectiveness and equity. *JAMA Netw Open*. 2021;4(4):e214622. doi:10.1001/jamanetworkopen.2021.4622

[80] Ansari MTJ, Al-Zahrani FA, Pandey D, Agrawal A. A fuzzy TOPSIS based analysis toward selection of effective security requirements engineering approach for trustworthy healthcare software development. *BMC Med Inform Decis Mak*. 2020;20(1):236. doi:10.1186/s12911-020-01209-8

[81] IBM. Adversarial robustness toolbox. GitHub. Accessed July 11, 2023. Retrieved from: https://github.com/Trusted-AI/adversarial-robustness-toolbox

[82] Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR Guiding Principles for scientific data management and stewardship [published correction appears in Sci Data. 2019 Mar 19;6(1):6]. *Sci Data*. 2016;3:160018. doi:10.1038/sdata.2016.18

[83] What is explainable AI (XAI)? IBM. Accessed July 11, 2023. Retrieved from: https://www.ibm.com/watson/explainable-ai

[84] Shaban-Nejad A, Michalowski M, Buckeridge DL. *Explainable AI in healthcare and medicine: building a culture of transparency and accountability.* Springer Cham; 2021. doi:10.1007/978-3-030-53352-6

[85] Arnold MH. Teasing out artificial intelligence in medicine: an ethical critique of artificial intelligence and machine learning in medicine. *J Bioeth Inq.* 2021;18:121–139. doi:10.1007/s11673-020-10080-1

[86] Centers for Medicare and Medicaid Services. CMS AI playbook, Version 2.0. CMS. October 2022. Accessed July 11, 2023. Retrieved from: https://ai.cms.gov/assets/CMS_AI_Playbook.pdf

[87] Ploug T, Holm S. The four dimensions of contestable AI diagnostics: a patient-centric approach to explainable AI. *Artif Intell Med*. 2020;107:101901. doi:10.1016/j.artmed.2020.101901

[88] Reporting guidelines. EQUATOR Network. Accessed April 14, 2023. Retrieved from: https://www.equator-network.org/reporting-guidelines/

[89] Gardiner LJ, Carrieri AP, Wilshaw J, Checkley S, Pyzer-Knapp EO, Krishna R. Using human in vitro transcriptome analysis to build trustworthy machine learning models for prediction of animal drug toxicity. *Sci Rep*. 2020;10(1):9522. doi:10.1038/s41598-020-66481-0

[90] Singh A. AI Explainability 360. IBM. August 2019. Accessed July 11, 2023. Retrieved from: https://www.ibm.com/opensource/open/projects/ai-explainability/

[91] Trusted AI. The Linux Foundation. Accessed July 11, 2023. Retrieved from: https://lfaidata.foundation/projects/trusted-ai/

[92] Marchesini K, Smith J, Everson J. Back to the future: What predictive decision support can learn from DeLoreans and The Big Short. Health IT Buzz. December 13, 2022. Accessed July 10, 2023. Retrieved from: https://www.healthit.gov/buzz-blog/health-innovation/back-to-the-future-what-predictive-decision-support-can-learn-from-deloreans-and-the-big-short

[93] Office of the Assistant Secretary for Planning and Evaluation. PCOR: Privacy and Security Blueprint, Legal Analysis and Ethics Framework for Data Use, & Use of Technology for Privacy. ASPE. Accessed September 5, 2023. Retrieved from: https://aspe.hhs.gov/pcor-privacy-security-blueprint-legal-analysis-ethics-framework-data-use-use-technology-privacy

[94] Hyatt Thorpe J, Cartwright-Smith L, Gray E, Mongeon M. Legal and ethical architecture for patient-centered outcomes research data. The Office of the National Coordinator for Health Information Technology. September 28, 2017. Accessed July 11, 2023. Retrieved from: https://www.healthit.gov/sites/default/files/page/2018-06/PCOR%20Architecture%20%28MERGE%29%20updated%20Appendix%20B.pdf

[95] Centers for Disease Control and Prevention. Legal and ethical framework to use Centers for Disease Control and Prevention data for patient-centered outcomes research. CDC. Accessed July 11, 2023. Retrieved from: https://aspe.hhs.gov/sites/default/files/private/pdf/259016/PCOR_Legal_508_2.pdf

[96] Office of the National Coordinator for Health Information Technology. What is FHIR®? ONC. Accessed July 11, 2023. Retrieved from: https://www.healthit.gov/sites/default/files/2019-08/ONCFHIRFSWhatIsFHIR.pdf

[97] Office of the National Coordinator for Health Information Technology. Training data for machine learning to enhance PCOR data infrastructure: implementation guidance. HealthIT.gov. Accessed July 11, 2023. Retrieved from: https://www.healthit.gov/sites/default/files/page/2021-09/ONC%20Training%20Data%20for%20ML_Implementation%20Guide.pdf

[98] Office of the National Coordinator for Health Information Technology. Synthetic health data generation to accelerate patient-centered outcomes research: Example modules and companion guides. HealthIT.gov. 2022. Accessed July 11, 2023. Retrieved from: https://www.healthit.gov/topic/scientific-initiatives/pcor/synthetic-health-data-generation-accelerate-patient-centered-outcomes

[99] Kennedy S. Researchers to propose framework to address federated learning challenges. Health IT Analytics. February 16, 2023. Accessed July 11, 2023. Retrieved from: https://healthitanalytics.com/news/researchers-to-propose-framework-to-address-federated-learning-challenges

[100] Bommasani R, Hudson DA, Adeli E, et al. On the opportunities and risks of foundation models. Center for Research on Foundation Models, Stanford Institute for Human-Centered Artificial Intelligence. 2021. doi.org/10.48550/arXiv.2108.07258

[101] Casusi J. What is a foundation model? An explainer for non-experts. Stanford University Human-Centered Artificial Intelligence. May 2023. Accessed July 27, 2023. Retrieved from: https://hai.stanford.edu/news/what-foundation-model-explainer-non-experts

[102] Moor M, Banerjee O, Abad ZSH. et al. Foundation models for generalist medical artificial intelligence. *Nature*. 2023;616: 259–265. doi.org/10.1038/s41586-023-05881-4

[103] Xu J, Glicksberg BS, Su C, Walker P, Bian J, Wang F. Federated learning for healthcare informatics. *J Healthc Inform Res*. 2021;5(1):1-19. doi:10.1007/s41666-020-00082-4

[104] Ng WY, Zhang S, Wang Z, et al. Updates in deep learning research in ophthalmology. *Clin Sci (Lond)*. 2021;135(20):2357-2376. doi:10.1042/CS20210207

[105] Tagde P, Tagde S, Bhattacharya T, et al. Blockchain and artificial intelligence technology in e-health. *Environ Sci Pollut Resh Int*. 2021;28(38):52810-52831. doi:10.1007/s11356-021-16223-0

[106] Collins GS, Dhiman P, Andaur Navarro CL, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open*. 2021;11(7):e048008. doi:10.1136/bmjopen-2020-048008

[107] Office of Research & Development, U.S. Department of Veterans Affairs. VA AI inventory. July 18, 2022. Accessed March 30, 2023. Retrieved from: https://www.research.va.gov/naii/ai-inventory.cfm

[108] Mayo Clinic research core facilities: about. Mayo Clinic. Accessed July 28, 2023. Retrieved from: https://www.mayo.edu/research/core-facilities/about

[109] National Cancer Institute. Core resources. Accessed July 27, 2023. Retrieved from: https://ostr.ccr.cancer.gov/resources/core

[110] Venkatesh V. Adoption and use of AI tools: a research agenda grounded in UTAUT. *Ann Op Res*. 2022;308(1-2), 641–652. doi:10.1007/s10479-020-03918-9

[111] Sounderajah V, Ashrafian H, Golub RM, et al. Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: the STARD-AI protocol. *BMJ Open*. 2021;11(6):e047709.. doi:10.1136/bmjopen-2020-047709

[112] Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK; SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med*. 2020;26(9):1364-1374. doi:10.1038/s41591-020-1034-x

[113] Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of Web-based and mobile health interventions. *J Med Internet Res*. 2011;13(4):e126. doi:10.2196/jmir.1923

[114] Des Jarlais DC, Lyles C, Crepaz N; TREND Group. Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: the TREND statement. *Am J Public Health.* 2004;94(3):361-366. doi:10.2105/ajph.94.3.361

[115] Rivera SC, Liu X, Chan AW, Denniston AK, Calvert MJ; SPIRIT-AI and CONSORT-AI Working Group. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI Extension. *BMJ*. 2020;370:m3210. Doi:10.1136/bmj.m3210

[116] Cambridge Dictionary. Algorithm. Accessed July 13, 2023. Retrieved from: https://dictionary.cambridge.org/us/dictionary/english/algorithm

[117] Pasick A. Artificial intelligence glossary: Neural networks and other terms explained. *The New York Times.* March 27, 2023. Accessed July 11, 2023. Retrieved from: https://www.nytimes.com/article/ai-artificial-intelligence-glossary.html

[118] Durán JM, Jongsma KR. Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. [published online ahead of print, 2021 Mar 18]. *J Med Ethics*. 2021;medethics-2020-106820. doi:10.1136/medethics-2020-106820

[119] Alfrink K, Keller I, Kortuem G, Doorn N. Contestable AI by design: Towards a framework. *Minds Mach*. 2022. doi:10.1007/s11023-022-09611-z

[120] Dilmegani C. What is AI augmentation? Techniques and examples in 2023. AI Multiple. Accessed July 11, 2023. Retrieved from: https://research.aimultiple.com/data-augmentation/

[121] Cambridge Dictionary. Dataset. Accessed July 11, 2023. Retrieved from: https://dictionary.cambridge.org/us/dictionary/english/dataset

[122] What is deep learning? IBM. Accessed July 11, 2023. Retrieved from: https://www.ibm.com/topics/deep-learning

[123] Nyuytiymbiy K. Parameters and hyperparameters in machine learning and deep learning. Towards Data Science. December 2020. Accessed August 2, 2023. Retrieved from: https://towardsdatascience.com/parameters-and-hyperparameters-aa609601a9ac

[124] National Institute of Standards and Technology. Metadata. NIST. Accessed July 11, 2023. Retrieved from: https://csrc.nist.gov/glossary/term/metadata

[125] University of North Carolina. Metadata for data management: A tutorial: Definition. Accessed July 11, 2023. Retrieved from: https://guides.lib.unc.edu/metadata/definition

[126] What is natural language processing? IBM. Accessed August 31, 2023. Retrieved from: https://www.ibm.com/topics/natural-language-processing

[127] What is overfitting? IBM. Accessed July 11, 2023. Retrieved from: https://www.ibm.com/topics/overfitting

[128] The Organization for Economic Cooperation and Development. Emerging privacy-enhancing technologies: Current regulatory and policy approaches. OECD. March 8, 2023. Accessed July 11, 2023. Retrieved from: https://www.oecd.org/publications/emerging-privacy-enhancing-technologies-bf121be4-en.htm

[129] AI Blindspot. Representative data. Accessed July 11, 2023. Retrieved from: https://aiblindspot.media.mit.edu/representative_data.html

[130] Five ways IBM is using synthetic data to improve AI models. IBM. Accessed July 11, 2023. Retrieved from: https://research.ibm.com/blog/synthetic-data-explained

[131] Barkved K. The difference between training data vs. test data in machine learning. ObviouslyAI. February 11, 2022. Accessed July 11, 2023. Retrieved from: https://www.obviously.ai/post/the-difference-between-training-data-vs-test-data-in-machine-learning

[132] What is underfitting? IBM. Accessed July 11, 2023. Retrieved from: https://www.ibm.com/topics/underfitting