CMSS
COUNCIL OF MEDICAL SPECIALTY SOCIETIES

ENCODING EQUITY
in Clinical Research & Practice
RETHINKING RACE IN CLINICAL ALGORITHMS

ENCODING EQUITY ALLIANCE: ARTIFICIAL INTELLIGENCE IMPLEMENTATION TOOLKIT

INTRODUCTION

The Encoding Equity Alliance is an alliance of organizations and individuals committed to driving change in clinical research and practice to advance health equity and optimize access and outcomes for all populations. Encoding Equity galvanizes collective action to amplify our impact, making change more quickly and comprehensively than would be possible for any one organization or group of constituencies on their own. Led by the Council of Medical Specialty Societies (CMSS), with support from the Doris Duke Foundation (DDF), the alliance engages and activates individuals and organizations across the medical, research, funding, publishing, and technology sectors.

The ultimate goal of these efforts is to advance health equity through a scientifically rigorous, evidence-driven, context-specific evaluation of the use and misuse of race and ethnicity in healthcare predictive algorithms and guidelines.

PREAMBLE

As the healthcare community increasingly integrates artificial intelligence (AI) into practice, the development of a cooperative comprehensive implementation framework is critical. AI embodies a dual potential: it can amplify human insight or entrench human error. Its promise lies in the capacity to analyze vast datasets and reveal patterns invisible to even the most experienced clinicians (Topol, 2019). Properly designed, AI serves as a "second intuition," offering diagnostic or treatment suggestions that seem almost prescient. Yet without safeguards, this "intuition" can reproduce racial, socioeconomic, and gender biases; discriminatory outcomes concealed beneath a veneer of objectivity (Obermeyer et al., 2019).

Paradoxically, the more confident we grow in AI's capabilities, the more vigilant we must become. Systems should be built to flag uncertainty, expose data limitations, and preserve human override authority (Amann et al., 2020). AI should complement human judgement rather than replace it, while avoiding the emergence of a new singular form of algorithmic hubris. This tension mirrors the age-old conflict between faith and fact: should we "trust" AI as one might trust divine wisdom, or should we trust AI after evidence-based evaluation? The answer lies in balance. Trust in AI must be conditionally earned through scientific reproducibility, transparency, auditability, and continuous validation (Babic et al., 2021). All must learn to engage AI critically, treating it as a collaborator rather than an oracle. In healthcare, AI should be evaluated on generalizability, accuracy, robustness, reliability, and clinical validity, ensuring that models perform well across diverse settings, maintain consistent predictive performance, and support safe, effective decision-making in real clinical environments without creating inequities.

While existing guidance addresses model validation, risk management, and software lifecycle practices, there remains limited explicit guidance on how health AI systems should demonstrate equitable performance across racialized and marginalized populations, nor clear accountability mechanisms when inequities are detected. In this regulatory vacuum, professional stakeholders bear an ethical responsibility to establish standards that safeguard equity and patient welfare. This requires an interconnected ecosystem of oversight, grounded in a collective commitment to non-maleficence and distributive justice. The risk, of course, is fragmentation: without a singular authority, evaluation standards could diverge, eroding public trust. Flexibility without coordination may yield inconsistency. To minimize fragmentation, any decentralized model must rest on unprecedented cooperation and mutual accountability among independent actors.

Yet even within structured systems, there lies a deeper challenge. Packets, frameworks, and guidelines, though well-intentioned, are vulnerable to devolving into tokenistic checklists, optimized for compliance rather than catalysts for persistent reflection. Once introduced, they often assume the status of key performance indicators (KPIs) that organizations strive to satisfy, and easily "game", rather than tools to continuously interrogate and improve underlying practice. Furthermore, a rapidly evolving technological landscape renders once appropriate solutions outdated. To prevent such outcomes, humility must be institutionalized. Decision-making regulatory processes should explicitly emphasize

doubt and dissent as mechanisms for refinement, not as threats to growth and authority. Multidisciplinary review boards, patient advocacy representation, and community ethics panels can democratize innovation and ensure accountability (Fricker, 2007). Epistemic injustice in healthcare AI arises not only when data misrepresents marginalized groups, but when their interpretive authority is systematically excluded from defining what counts as evidence, harm, or success.

AI, whether savior or saboteur, must be deployed with critical self-awareness, reinforcing the best of our instincts while countering our worst biases. It is therefore crucial that institutions measure the metrics that AI is intending to solve. If those metrics show no meaningful progress, the appropriate response is not to recalibrate the targets, but to return to the drawing board and reiterate. Authentic accountability demands continuous recalibration of purpose, not mere procedural adherence.

In the absence of centralized regulation, this framework seeks to promote a shared professional commitment: that technological innovation in healthcare must enhance accuracy and efficiency in the service of justice, transparency, and trust.

MISSION STATEMENT:

This toolkit seeks to advance the development and deployment of healthcare artificial intelligence systems that improve health outcomes and are scientifically rigorous, ethically grounded, and equitably aligned by eliminating the inappropriate use of race as a proxy variable. It aims to replace race-based modeling with approaches that more accurately reflect the biological, environmental, and structural determinants of health, thereby improving clinical decision-making, reducing disparities, and strengthening trust among patients, providers, and institutions. Through transparency, accountability, and multidisciplinary collaboration, the document aspires to set a standard for responsible innovation that serves both individual patients and the broader public good.

GUIDING VALUES AND OBJECTIVES:

A rigorous framework for the exclusion of race in healthcare artificial intelligence must begin from a clear ethical and scientific premise: that race, as commonly encoded in clinical and administrative datasets, is an imprecise social construct that often functions as a proxy for structural inequities rather than a biologically valid variable (Vyas et al., 2020; González-Burchard et al., 2003). The uncritical inclusion of race in algorithmic systems risks perpetuating historical biases, reinforcing unequal treatment pathways, and obscuring the true drivers of health disparities (Obermeyer et al., 2019). A responsible toolkit must therefore articulate values and goals that improve precision and shift the focus from race-based adjustment toward equity-driven, causally grounded, and clinically meaningful modeling.

Guiding Objectives

Eliminate the routine use of race as a default variable in predictive models unless its inclusion can be justified by clear, evidence-based, and ethically defensible reasoning (Vyas et al., 2020).

Advance the development and integration of alternative variables that more accurately capture the drivers of health outcomes, including socioeconomic status, environmental exposures, and healthcare access (Centers for Disease Control and Prevention, 2023).

Standardize evaluation protocols that measure both overall performance and subgroup equity, ensuring that improvements in aggregate accuracy do not mask harm to vulnerable populations (Rajkomar et al., 2018).

Create interoperable documentation standards that allow systems to be scrutinized, compared, and improved across institutions (Mitchell et al., 2019; Gebru et al., 2018).

Foster a culture of continuous learning in which developers and healthcare organizations remain responsive to new evidence, stakeholder feedback, and evolving ethical norms.

Guiding Values

EQUITY Equity is a core operational principle embedded throughout the AI lifecycle, from problem formulation and data collection to deployment and monitoring. Models should undergo disaggregated performance evaluation to ensure that removing race does not worsen outcomes for historically marginalized groups. Equity is achieved not by ignoring differences but by identifying and addressing the causes of unequal outcomes. Developers and institutions must demonstrate that their systems reduce, rather than perpetuate, existing disparities.

TRANSPARENCY & INTERPRETABILITY
Transparency and interpretability are essential in healthcare AI because algorithmic decisions affect patient outcomes, trust, and regulatory compliance. Systems should clearly document variable selection, model assumptions, and limitations. When race is excluded, stakeholders must understand what variables replace it and the rationale for doing so. Transparency supports informed oversight by clinicians, administrators, and patients, while interpretability ensures that AI remains subject to clinical judgment rather than functioning as an opaque decision-maker.

SCIENTIFIC INTEGRITY Models should prioritize variables with clear physiological or environmental relevance rather than race. This requires scrutinizing inputs: if race appears predictive, the question is why. Predictive value often reflects unmeasured factors like healthcare access, environmental exposures, socioeconomic conditions, or structural discrimination. Race should be replaced with direct measures of these underlying determinants whenever possible, or, when such measures are unavailable, uncertainty should be explicitly acknowledged instead of relying on proxy variables.

PATIENT-CENTEREDNESS Excluding race from AI is not merely a technical adjustment but a commitment to respecting individuals as more than categorical abstractions. Systems should be designed to reflect the lived realities of patients, incorporating social determinants of health in ways that are context-sensitive and non-stigmatizing. Engaging patients and communities in the design and evaluation of these systems can help ensure that the resulting tools align with their needs, values, and expectations and foster trust.

ACCOUNTABILITY Must extend beyond technical performance to ethical impact. Institutions deploying AI systems should establish governance structures that include multidisciplinary oversight, incorporating expertise from clinicians, data scientists, ethicists, and community representatives. Continuous auditing mechanisms should be implemented to detect emergent biases, performance drift, and unintended consequences. Accountability also implies the willingness to revise or withdraw systems that fail to meet equity and safety standards, even when they perform well on conventional metrics.

AI LIFECYCLE:

As AI becomes increasingly embedded in healthcare delivery, there is a growing need for evidence-informed governance frameworks that ensure these technologies advance equity rather than reproduce or exacerbate existing biases. This toolkit provides structured guidance to operationalize equity across the AI life cycle by explicitly linking six primary drivers to each stage of AI development, implementation, and oversight.

A critical but often overlooked stage in this life cycle is problem formulation. Decisions about which problems AI is intended to address, and whose definitions of "success" are prioritized fundamentally shape downstream equity outcomes, often to a greater extent than technical design choices. Too often efforts to improve the value of health care largely ignore improving health outcomes and concentrate on reducing costs and increasing shareholder profit, a growing concern as private equity increases its share of the healthcare marketplace (Kannan et al, 2025 and Kannan et al., 2023). Prior to initiating development, organizations should formally document: (1) the clinical or operational problem motivating AI use; (2) the rationale for selecting an algorithmic approach over heuristic alternatives; and (3) the stakeholders who may benefit from or be burdened by automation. Failure to interrogate these foundational assumptions risks embedding inequity before development begins.

Each primary driver within the toolkit is accompanied by actionable interventions, implementation strategies, and measurable evaluation metrics. Data creation is guided by principles of *Data Equity and Representation*, emphasizing the collection of high-quality, complete, and clinically relevant data that reflect health-relevant diversity while avoiding inappropriate proxies, including the misuse of race. Data acquisition is aligned with *Sound Causal Inference*, prioritizing careful dataset integration and the use of causal methods to ensure that model inputs reflect clinically meaningful relationships rather than spurious correlations.

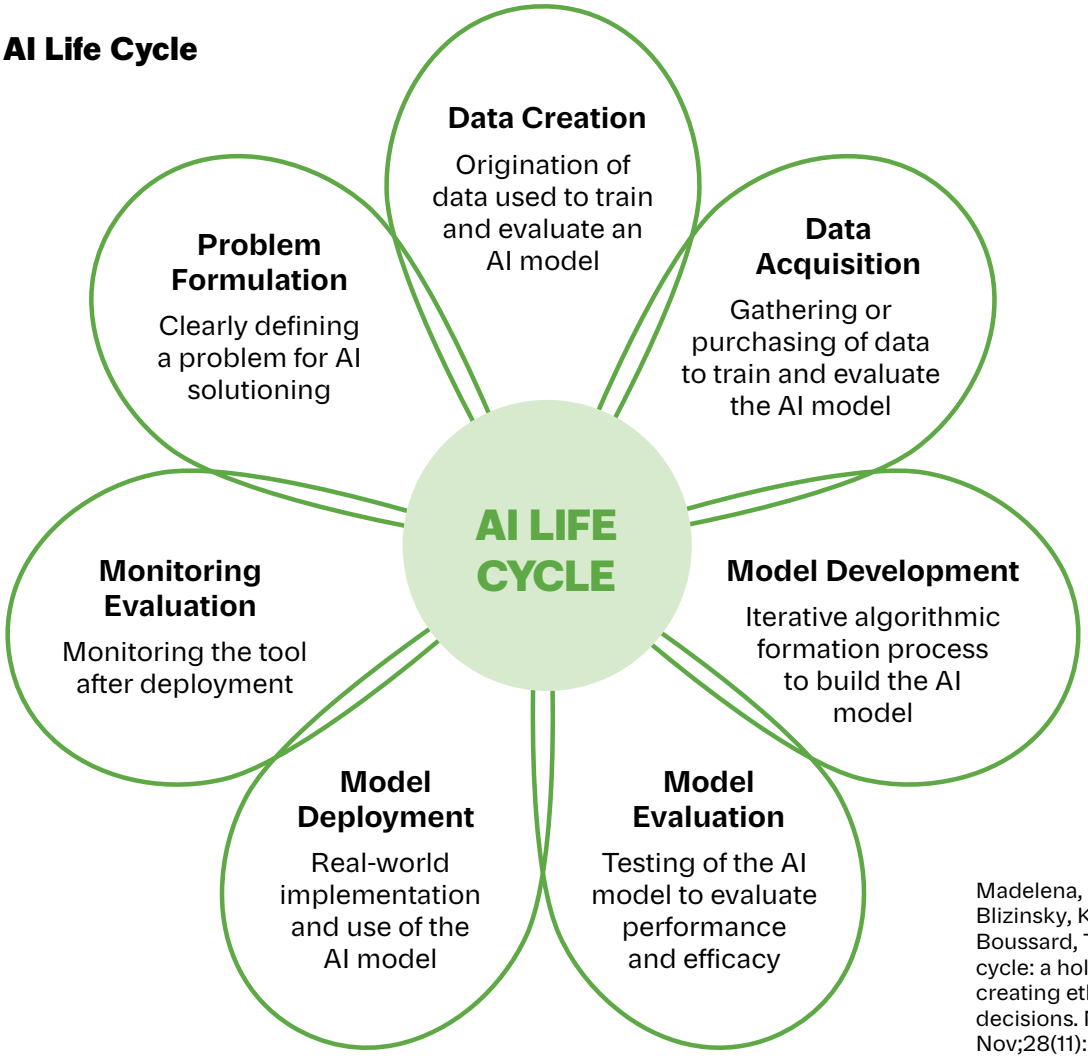
Model development is supported by *Model Transparency and Explainability*, which promote clear documentation of design decisions, feature selection, and ground-truth labeling to enhance interpretability and reduce the risk of embedded bias. Model evaluation is anchored in *Algorithmic Accountability*, requiring rigorous performance testing, fairness assessments, and subgroup analyses to identify and address inequities. *Inclusive Governance and Stakeholder Engagement* are integrated into the evaluation phase to provide oversight, incorporate diverse perspectives, and guide responsible decision-making.

Model deployment is aligned with *Required Ethical Reviews*, incorporating structured approval processes, post-deployment monitoring, and safeguards to detect unintended consequences and ensure that outputs remain grounded in valid clinical indicators. Together, these drivers form a coherent, lifecycle-oriented framework for proactively identifying, mitigating, and continuously monitoring bias in healthcare AI systems.

Effective AI governance must be lifecycle-oriented, spanning problem framing, development, validation, deployment, monitoring, and decommissioning. However, sustained oversight depends on the availability of appropriate technical and institutional infrastructure. In practice, many healthcare organizations-particularly under-resourced hospitals and safety-net systems-lack the data access, technical capacity, contractual leverage, or staffing required for continuous auditing, monitoring, and model retirement. AI systems designed for high-resource environments or centralized expertise risk concentrating accountability upstream while limiting local oversight.

Accordingly, the primary drivers outlined in this toolkit should be understood not as discrete requirements, but as interdependent safeguards that must be maintained throughout the AI life cycle. Their effectiveness depends on organizational infrastructure that supports ongoing review, intervention, and, when necessary, decommissioning to ensure that AI tools deliver meaningful and equitable benefits in healthcare.

AI Life Cycle



Madelena, Y Ng., Supriya, K., Blizinsky, K., & Hernandez-Boussard, T. (2023). The AI life cycle: a holistic approach to creating ethical AI for health decisions. Nat Med. 2022 Nov;28(11):2247–2249. doi: 10.1038/s41591-022-01993-y

DISTRIBUTED OPEN JUSTICE OVERSIGHT

OPERATIONALIZING EQUITABLE AI GOVERNANCE WITH DOJO

The toolkit defines six drivers of equitable AI governance, each presupposing institutional capacity for technical evaluation. In practice, this capacity is limited:

- *Data equity* requires auditing training datasets
- *Algorithmic accountability* requires tracing assumption lineage
- *Causal inference* requires subgroup-based counterfactual testing

Most hospitals lack the expertise and infrastructure to perform these functions. Vendors are not incentivized to conduct them independently, and regulatory requirements remain insufficient.

DOJO: DISTRIBUTED OPEN JUSTICE OVERSIGHT

DOJO is an open-source platform designed to enable adversarial evaluation of health AI systems across institutions, regardless of technical capacity. It operationalizes the toolkit's drivers through modular evaluation agents, each targeting specific failure modes:

- *Shortcut Detection Agent:* Identifies reliance on spurious signals (e.g., imaging artifacts, institutional markers) rather than clinical pathology.
- *EquiFlow Agent:* Tracks cohort composition across inclusion/exclusion steps during study design.
- *Measurement Equity Agent:* Assesses consistency in clinical documentation practices across patient subgroups.
- *Care Phenotypes Agent:* Evaluates model performance across treatment patterns rather than demographic proxies.

System Design and Accessibility

DOJO is built for low-barrier adoption:

- Containerized and deployable via a single command
- Operates entirely within institutional firewalls (no external API calls)
- Uses a standardized orchestration protocol (MCP) for extensibility

Community-contributed agents can be integrated through an open submission process. All agents undergo automated compliance and reproducibility checks; validated agents are released for general use, while others remain flagged as experimental.

Functional Impact

DOJO redistributes evaluation responsibility from institutions to tooling:

- Enables non-specialist settings to conduct rigorous model audits
- Produces structured reports on shortcut features, subgroup disparities, and measurement bias
- Logs all evaluations, parameters, and outputs to generate transparent audit trails

Stakeholders (i.e. ethicists, administrators, clinicians, patients, and community health workers) can contribute adversarial test cases reflecting local contexts and edge conditions often absent from standard benchmarks.

Role Within the Toolkit

DOJO does not replace the governance framework; it renders it actionable:

- *Bias tracing ("bias archaeology"):* operationalized via shortcut detection and EquiFlow agents
- *Disaggregated evaluation:* implemented through care phenotypes analysis
- *Deployment oversight:* supported by reproducible, auditable evidence

By converting governance principles into executable processes, DOJO enables enforcement in under-resourced settings where such standards are otherwise unattainable.

Ecosystem Integration

DOJO is developed within an established global research network, leveraging existing data infrastructure and collaborative capacity. Its primary function is to make the toolkit's six drivers measurable, testable, and enforceable in real-world clinical environments.

PURPOSE

This toolkit is not intended to replace regulatory requirements, institutional review processes, or clinical judgment. Rather, it functions as an operational governance resource that translates ethical commitments to equity, accountability, and transparency into concrete practices that can be embedded within existing AI development, and oversight workflows. Its purpose is to support more rigorous questioning, stronger safeguards, and clearer decision-making, rather than to prescribe specific technical solutions.

Although this toolkit is not designed to provide detailed guidance on general-purpose AI applications, including publicly accessible platforms such as publicly accessible health AI platforms, we acknowledge their growing role in health-related decision-making and underscore the importance of continued scrutiny regarding their potential impact on racial equity.

How to use this toolkit:

- As a quality improvement tool during model development and validation
- As an evaluation and procurement guide for health systems
- As a governance framework for oversight committees
- As an educational resource for clinicians, trainees, and researchers

Use Case

A midwestern 600-bed academic medical center with an AI governance office, in-house data scientists, and an equity advisory board that can block procurement decides to buy a commercial sepsis prediction tool. The toolkit's six drivers map onto the infrastructure the hospital already has. The data team audits the training set for demographic gaps and finds that the model's cohort underrepresents the hospital's large Haitian and Cape Verdean communities. The causal inference review reveals that the marker of acuity is measured as "number of prior ED visits," which tracks insurance status. The transparency driver forces the vendor to produce model cards documenting feature provenance, ground-truth labeling, and subgroup performance. When the ethics committee finds a 12-percentage-point disparity in false-negative rates between Black patients and nonblack patients, it halts deployment until the vendor remediates. The toolkit works here because the institution can enforce all its drivers and has the staff and budget to follow through on post-deployment monitoring.

The same tool arrives at an Appalachian 30-bed rural safety-net hospital with one intensivist and no data scientist. Its patients are disproportionately uninsured, Indigenous, or from communities whose encounters never appeared in the training data. Every driver in the toolkit remains ethically valid. None of them can be operationalized. There is no equity advisory board, no capacity for causal feature review, no leverage to demand disaggregated performance data from a vendor whose sales team calls the model "FDA-reviewed." The parable of the sepsis algorithm that couldn't listen surfaces here in concrete form: a governance framework designed for the resourced is most needed where resources are thinnest. The response should be to change who bears the burden. Contracts should require vendors to demonstrate equitable performance in the purchaser's population before deployment, with the cost of that evaluation borne by the seller. Open-source audit platforms, operated

by communities of practice rather than by corporate compliance offices, could pool evaluation capacity across small hospitals that cannot build it alone. Regional coalitions could let a safety-net hospital in Appalachia and a critical-access hospital in the Delta share interpretive authority over what counts as harm. The toolkit's own preamble warns against frameworks that reward procedural adherence over authentic accountability. If the governance architecture cannot reach the places where patients are most exposed, it protects only the institutions that least need protection.

Key Tensions: Procurement vs Contractual Mechanisms in AI Acquisition

Procurement mechanisms structure the competitive selection and evaluation of third-party AI suppliers prior to award, while contractual mechanisms define the allocation of rights, risks, and responsibilities governing the deployment and lifecycle management of AI systems post-award; together, they form an integrated framework for ensuring value, accountability, and regulatory compliance in AI health tool(s) acquisition.

This toolkit focuses exclusively on contractual mechanisms governing the acquisition and deployment of AI systems and does not address procurement processes related to vendor selection or tendering.

This toolkit applies an implementation science and quality improvement framework to support the development, deployment, and evaluation of equitable AI tools in healthcare settings. The accompanying tables guide teams in operationalizing bias-mitigating practices across the AI lifecycle. The tables identify key drivers of equity, highlight stages where bias may emerge or be reinforced, and outline corresponding change concepts, interventions, and actionable change. Each table also includes evaluation metrics to support continuous monitoring, accountability, and measurement of outcomes.

KEY FACTOR: Biased Training Data – Historical inequities embedded in datasets that yield unequal model performance across racial and ethnic groups.

PRIMARY DRIVER 1: DATA EQUITY & REPRESENTATION	
Change Concept	**Ensure training data reflects the diversity of the population served**
Interventions	Build and use demographically diverse datasets • Meaningful representation across race/ethnicity, gender, age, income, education, geographic region, language, disability status, language, nationality or culture Sample size for each group is large enough for analysis and the distribution approximates the target population or intentionally balances groups
	Audit data sources regularly for imbalances
	Partner with community health systems for inclusive data collection
Change Idea(s)	Create data dashboards that monitor representation
	Implement third-party data bias audits before model training
	Develop protocols for rejecting datasets with known biases
	More open-source data sets
Evaluation Metrics	Minimum percentage of training data representing each demographic with descriptions (applicable to patient population)
	Frequency of dataset bias audits (programs, vendors, screening vs diagnostic)

Evaluation Metrics cont.	Data completeness across key demographic fields

Synthetic Data

When paired with robust fairness evaluation and grounded in diverse real-world data, synthetic datasets can support innovation while promoting privacy, inclusivity, and more equitable model development. Because synthetic data sets inherit the statistical patterns, and therefore the biases, from its source data, it cannot compensate for underlying gaps or misrepresentation in the real-world population. Developers must critically assess whether the underlying data meaningfully represents racialized and marginalized groups, document potential biases that may persist or be amplified in generated datasets, and ensure synthetic outputs do not obscure nor reproduce existing health inequities. Transparent reporting of data provenance, generation techniques, demographic coverage, and known limitations is essential so that stakeholders understand appropriate use cases and avoid overreliance on synthetic data generated outputs for high-stakes clinical or policy decisions. Synthetic data should not be used as a substitute for real-world validation in high-stakes clinical or policy decisions where inequitable impacts could plausibly result in harm.

KEY FACTOR: Legacy Assumptions – Outdated or racialized paradigms informing model design and feature selection.

PRIMARY DRIVER 2: ALGORITHMIC ACCOUNTABILITY	
Change Concept	**Reassess legacy assumptions and technical choices that embed structural bias**
Interventions	Reevaluate algorithmic logic derived from outdated or biased clinical knowledge
	Incorporate domain experts to review the clinical relevance of output through an equity lens
	Update models using more recent and inclusive evidence; include versioning history

Interventions cont.	Conduct periodic assumption lineage reviews ("bias archaeology") to trace how specific features, proxies, or clinical heuristics entered the model, including historical practices, legacy guidelines, or prior discriminatory norms that may still shape algorithmic behavior. • Note: Assumptions may include, but are not limited to, choices of outcome definitions, proxy variables, clinical heuristics embedded in features, labeling practices, and implicit causal interpretations derived from correlational data.
Change Ideas	Host legacy assumption reviews during model updates
	Use literature reviews to identify biased clinical heuristics
	Build model cards detailing implications of assumptions
Evaluation Metrics	Number of assumptions reviewed per cycle/iteration • Importance of assumptions needs to be weighted
	Number of clinical practices flagged as biased
	Proportion of models with updated equity-focused documentation
	Performance across diverse subgroups/phenotypes (race, ethnicity, SDoH, sexual identity/orientation, populations with least care etc.)

KEY FACTOR: **Opaque Models** – Limited transparency in algorithmic development, validation, and decision-making that obscures bias detection.

PRIMARY DRIVER 3: MODEL TRANSPARENCY & EXPLAINABILITY	
Change Concept	**Understand data points and intended use of model**
Intervention	Adopt equity-centered explainability practices that not only clarify how predictions are generated, but explicitly surface whether and how model behavior differs across racialized and marginalized groups. Share interpretable model outputs with clinicians and patients.
Change Idea	Visualize disparities in feature importance by social determinants (i.e. race and ethnicity). Publish simplified decision logic for non-technical stakeholders. • Explainability plays a critical role in allocating accountability within AI-enabled healthcare systems. While "human-in-the-loop" models are often invoked to ensure safety, responsibility in practice is distributed across multiple actors, including developers, healthcare institutions, vendors, regulators, and clinicians. Without clarity on how decisions are generated and constrained, accountability risks being implicitly shifted onto frontline clinicians by default. Explainability artifacts should, therefore, be designed not only to support clinical decision-making, but to make visible the assumptions, limitations, and constraints imposed by upstream design, data, and institutional deployment choices—so that responsibility is appropriately shared rather than silently transferred. In the absence of clear accountability mapping, explainability requirements can paradoxically increase moral burden on clinicians while shielding upstream actors from scrutiny.

Evaluation Metrics	Team Card: who created the algorithm and what types of engagement were used when developing the model Best practice: include philosophers, social scientists and clinicians
	Documented provenance: documented history of data, outlining its origin, transformations and usage (i.e. peer-reviewed, high-quality guidelines) with citations
	Documented auditability: all actions related to data are recorded and verifiable, creating an audit trail
	Publish simplified decision logic
	Dashboards stratified by demographic group, alerts for performance divergence, and counterfactual explanations that test sensitivity to social identity variables
	Explainability requirements should be proportionate to stakeholder needs: clinicians require clinically interpretable reasoning aligned with workflows, patients require accessible explanations of how AI informs care, and regulators require traceable documentation supporting safety and equity claims
	Organizations should document named accountable party for each life cycle stage (design, validation, deployment, monitoring, override), and ensure that corresponding explainability mechanisms exist for each role

KEY FACTOR: Exclusion of Voices – Underrepresentation of diverse patients, practitioners, and communities in AI design and governance.

PRIMARY DRIVER 4: INCLUSIVE GOVERNANCE & STAKEHOLDER ENGAGEMENT	
Change Concept	**Involve diverse voices in AI development and oversight**
Interventions	Involve sub-groups in model co-design
	Formalize inclusion of marginalized voices in governance structures • Formalize shared governance with marginalized communities in decision-making structure, or; • Center marginalized communities as decision-makers within governance structures, or; • Redistribute decision-making authority to include marginalized communities; and, a diverse group of healthcare professionals (who use AI at the point of care) as governing partners *Engagement alone does not constitute equity. Equity requires that stakeholder input has the capacity to meaningfully alter design decisions, delay deployment, or trigger de-implementation.*
	Use participatory methods to elicit lived experiences from a diverse and representative group of stakeholders including health care professionals and community members
Change Ideas	Create equity advisory boards with decision authority
	Host co-design workshops with community members and health care professionals who are potential users of AI at the point of care
	Equity impact statements in development cycles

Change Ideas cont.	Basic aspects of AI in education (i.e. PhD/PhD level students, medical students, residents/fellows, APP's, nurses, pharmacists, etc.)
Evaluation Metrics	Diversity of stakeholder groups consulted
	Meaningful engagement extends beyond consultation; governance structures should explicitly document how stakeholder input informs decisions, including instances where recommendations lead to model modification, delayed deployment, or rejection
	Stakeholder satisfaction with involvement process

KEY FACTOR: Correlation Misuse – Mistaking statistical association for causation, perpetuating stereotypes and ignoring social determinants of health (SDoH).

PRIMARY DRIVER 5: SOUND CAUSAL INFERENCE	
Change Concept	**Move beyond correlation to identify fair and valid predictors of outcomes**
Interventions	Use causal modeling to distinguish correlation from causation. Failure to distinguish correlational prediction from causal inference risks inappropriate downstream use of AI outputs, particularly when models inform resource allocation, risk stratification, or policy decisions affecting marginalized populations
	Train teams to critically assess feature inclusion (i.e. biological/genetic underpinnings)
	Engage a third-party entity to evaluate the model and vendor
	Validate models across stratified populations

Change Ideas	Use models for prediction, NOT interventions (medical knowledge for intervention)
	Clearly state the goal/purpose of the algorithm: • Specify if the AI is correlational (identifying statistical associations in data) or causal (identifying cause-and-effect relationships). This distinction determines the required data, assumptions, validation strategies, and the kinds of clinical decisions AI can appropriately support. AI documentation should explicitly state this designation so that end users understand the model's intended capabilities, limitations, and appropriate clinical use
	Prohibit the use of variables with unexplainable disparities
	Apply casual inference methods for bias detection
	Conduct counterfactual testing across groups
Evaluation Metrics	Percentage of models using causal methods
	Percentage of disparities in false positives/negatives
	Number of features removed due to non-causal correlation with race

KEY FACTOR: **Commercial Speed** – Market imperatives that prioritize rapid deployment over rigorous ethical and fairness evaluation.

PRIMARY DRIVER 6: REQUIRED ETHICAL REVIEWS	
Change Concept	**Balance innovation with rigorous ethical review in time-sensitive environments**
Interventions	Integrate equity and ethics checks in all stages of the AI lifecycle
	Empower internal ethics reviewers with decision-blocking authority
	Require pre-deployment fairness testing
	A legitimate outcome of ethical and equity review is the decision not to build, deploy, or continue an AI system
Change Ideas	Use AI ethics checklists at each stage of product development
	Establish model go/no go gates based on equity audits
	Mandate post launch audits to monitor model drift, inequitable performance, unintended consequences, and safety issues, with mechanisms for recalibration, rollback, and reporting of adverse events. Integrate ethics and equity review into development workflows through parallel “ethics sprints,” in which fairness testing, stakeholder review, and bias assessment occur alongside technical development rather than as post-hoc checkpoints.
	Implement equity-focused red-team exercises in which interdisciplinary teams actively probe models for disparate impacts, proxy variables, and failure modes affecting marginalized populations prior to deployment

Evaluation Metrics	Number of ethics checklists conducted
	Number of post-launch audits conducted with proportionate corrective action (inclusive of model refinement to decommissioning)

GLOSSARY OF TERMS

Algorithm
A finite sequence of well-defined computational steps used to transform inputs into outputs to solve a problem or perform a task.

Algorithmic Bias
Systematic and unfair discrimination produced by algorithms when predictions disproportionately disadvantage certain populations due to biased training data or model design.

Algorithmic Fairness
The principle that AI and machine learning systems should distribute predictions or decisions equitably across demographic groups without systematic disadvantage.

Artificial Intelligence (AI)
A field of computer science focused on designing systems capable of performing tasks that typically require human intelligence, such as reasoning, learning, and perception.

Artificial Intelligence in Healthcare
The application of AI methods (including machine learning, natural language processing, and predictive analytics) to medical data to support clinical decision making, improve diagnostics, and inform treatments.

Causal Inference
A set of statistical and methodological approaches used to determine whether relationships between variables represent cause and effect rather than simple correlation.

Clinical Decision-Making
The cognitive and analytical process used by healthcare professionals to evaluate patient information and determine appropriate diagnoses, treatments, and management strategies.

Clinical Decision Support Systems (CDSS)
Health information systems that provide clinicians with patient specific analysis or recommendations to support healthcare decisions.

Clincial Guidelines
Systematically developed statements intended to assist practitioner and patient decisions about appropriate healthcare for specific clinical problems.

Clinical Prediction Models
Statistical or machine learning models that estimate the probability of clinical outcomes based on patient characteristics and clinical data.

Data (Structured & Unstructured)
Recorded observations representing information about variables or phenomena that can be analyzed. Structured data follow predefined formats (e.g., tables), whereas unstructured data lack fixed schema (e.g., clinical text).

Dataset Shift / Model Drift
Changes in the statistical properties of input data over time that may reduce the accuracy or reliability of predictive models.

Deep Learning
A class of machine learning methods that use multilayer neural networks to learn hierarchical representations of data.

Disparities (Health Disparities)
Preventable differences in health outcomes or healthcare access experienced by socially disadvantaged populations.

Electronic Health Record (EHR)
A digital version of a patient's medical and treatment history designed for use across healthcare settings.

Explainability
The extent to which the internal processes and decision logic of an AI system can be understood by humans.

Generative Artificial Intelligence
AI systems capable of producing new, original content (e.g., text or images) learned from large datasets.

Guidelines
Systematically developed recommendations intended to guide decisions by providing evidence based advice.

Health Equity
The attainment of the highest possible standard of health for all people by addressing avoidable inequalities.

Health Informatics
An interdisciplinary field that applies information and computer science to improve healthcare outcomes.

Large Language Models (LLMs)
Advanced deep learning models trained on extensive text corpora to perform language tasks such as generation and summarization.

Machine Learning
Computer algorithms that allow systems to learn patterns from data and improve performance on tasks without explicit programming.

Model Validation (Clinical AI)
The assessment of an AI model's performance on independent data to determine generalizability and reliability.

Natural Language Processing (NLP)
A field of artificial intelligence focused on enabling computers to understand, interpret, and generate human language.

Precision Medicine
A healthcare approach that tailors medical decisions and interventions to individual patient characteristics (e.g., genetics, environment).

Predictive Analytics
The use of statistical or machine learning models to forecast future health outcomes based on historical data.

Responsible AI
Frameworks and practices that emphasize fairness, accountability, transparency, privacy, and ethical considerations in AI system development and use.

Transparency
Openness and clarity in describing AI system data sources, model structure, and decision processes to stakeholders.

Workflow Integration (AI in Healthcare)
The incorporation of AI systems into clinical and operational healthcare workflows to enhance efficiency and decision making.

REFERENCES

American Medical Association. (2018). *Augmented intelligence in health care*. American Medical Association.

Ashley, E. A. (2016). Towards precision medicine. *Nature Reviews Genetics*, 17(9), 507–522

Amann, J., Blasimme, A., Vayena, E., Frey, D., & Madai, V. I. (2020). Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Medical Informatics and Decision Making, 20*(1), 310. https://doi.org/10.1186/s12911-020-01332-6

Babic, B., Gerke, S., Evgeniou, T., & Cohen, I. G. (2021). Beware explanations from AI in health care. Science, 373(6552), 284–286. https://doi.org/10.1126/science.abg1834

Bates, D. W., Saria, S., Ohno Machado, L., Shah, A., & Escobar, G. (2014). Big data in healthcare: Using analytics to identify and manage high risk and high cost patients. *Health Affairs, 33*(7), 1123–1131

Berner, E. S. (2007). *Clinical decision support systems: Theory and practice* (2nd ed.). Springer

Bielick CG, Awwad A, Ellen J, Jalilian L, McCoy LG, Mishra V, et al. (2026) Moving beyond the benchmarks: Five foundational principles for meaningful AI evaluation in healthcare. PLOS Digit Health 5(5): e0001115. https://doi.org/10.1371/journal.pdig.0001115

Braveman, P. (2006). Health disparities and health equity: Concepts and measurement. *Annual Review of Public Health*, 27, 167–194

Braveman, P., & Gruskin, S. (2003). Defining equity in health. *Journal of Epidemiology & Community Health*, 57(4), 254–258

Braveman, P., Kumanyika, S., Fielding, J., LaVeist, T., Borrell, L., Manderscheid, R., & Troutman, A. (2011). Health disparities and health equity: The issue is justice. *American Journal of Public Health, 101*(S1), S149–S155. https://doi.org/10.2105/AJPH.2010.300062

Celi, L. A., Cellini, J., Charpignon, M. L., et al. (2022). Sources of bias in artificial intelligence that perpetuate healthcare disparities—A global review. PLOS Digital Health 1(3): e0000022. https://doi.org/10.1371/journal.pdig.0000022

Centers for Disease Control and Prevention. (2023). *Social determinants of health: Know what affects health*. https://www.cdc.gov/socialdeterminants

Collins, F. S., & Varmus, H. (2015). A new initiative on precision medicine. *New England Journal of Medicine*, 372(9), 793–795

Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2022). *Introduction to algorithms* (4th ed.). MIT Press

Dahabreh IJ, Bibbins-Domingo K. Causal Inference About the Effects of Interventions From Observational Studies in Medical Journals. JAMA. 2024 Jun 4;331(21):1845-1853. doi: 10.1001/jama.2024.7741. PMID: 38722735

Dullabh P, Dhopeshwarkar R, Leaphart D, et al. Trustworthy Artificial Intelligence (TAI) for Patient-Centered Outcomes Research (PCOR): Report [Internet]. Washington (DC): Office of the Assistant Secretary for Planning and Evaluation (ASPE); 2023 Sep. https://www.ncbi.nlm.nih.gov/books/NBK605645/

Eccles, M., & Mason, J. (2001). How to develop cost conscious guidelines. *Health Technology Assessment, 5*(16), 1–69

Elstein, A. S., & Schwartz, A. (2002). Clinical problem solving and diagnostic decision making: Selective review of the cognitive literature. *BMJ, 324*(7339), 729–732

Emmert-Streib, F., Yli-Harja, O., & Dehmer, M. (2020). Explainable artificial intelligence and machine learning: A reality rooted perspective. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(5), e1368. https://doi.org/10.1002/widm.1368

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., & Dean, J. (2019). A guide to deep learning in healthcare. *Nature Medicine*, 25, 24–29. https://doi.org/10.1038/s41591-018-0316-z

Flanagin A, Lewis RJ, Muth CC, Curfman G. What Does the Proposed Causal Inference Framework for Observational Studies Mean for JAMA and the JAMA Network Journals? JAMA. 2024 Jun 4;331(21):1812-1813. doi: 10.1001/jama.2024.8107. PMID: 38722708

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... Schafer, B. (2018). AI4People—An ethical framework for a good AI society. *Minds and Machines, 28*(4), 689–707

Fricker, M. (2007). *Epistemic injustice: Power and the ethics of knowing*. Oxford University Press.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. Communications of the ACM 64 (12), 86-92. https://doi.org/10.1145/3458723

González Burchard, E., Ziv, E., Coyle, N., Gomez, S. L., Tang, H., Karter, A. J., Mountain, J. L., Pérez-Stable, E. J., Sheppard, D., & Risch, N. (2003). The importance of race and ethnic background in biomedical research and clinical practice. *New England Journal of Medicine, 348*(12), 1170–1175. https://doi.org/10.1056/NEJMsb025007

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press

Goodman, S. N., Goel, S., & Cullen, M. R. (2018). Machine learning, health disparities, and causal reasoning. *Annals of Internal Medicine, 169*(12), 883–884. https://doi.org/10.7326/M18-3297

Hernán, M. A., & Robins, J. M. (2020). *Causal inference: What if*. Chapman & Hall/CRC

Hunink, M., Weinstein, M., Wittenberg, E., Drummond, M., Pliskin, J., Wong, J., & Glasziou, P. (2014). *Decision making in health and medicine* (2nd ed.). Cambridge University Press

Institute of Medicine. (2011). *Clinical practice guidelines we can trust*. National Academies Press. https://doi.org/10.17226/13058

Institute of Medicine. (2001). *Crossing the quality chasm: A new health system for the 21st century*. National Academies Press.

Institute of Medicine. (2003). *Unequal Treatment: Confronting racial and ethnic disparities in healthcare*. National Academies Press

International Organization for Standardization. (2022). *ISO/IEC 22989: Artificial intelligence—Concepts and terminology*. ISO

Jagadish, H. V., et al. (2014). Big data and its technical challenges. *Communications of the ACM, 57*(7), 86–94

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science, 349*(6245), 255–260

Jurafsky, D., & Martin, J. H. (2023). *Speech and language processing* (3rd ed.). Pearson

Kannan, S., Bruch J.D., Song, Z. (2023). Changes in Hospital Adverse Events and Patient Outcomes Associated with Private Equity Acquisition. *Journal of the American Medical Association*, 26; 330 (24), 2365-2375

Kannan, S., Bruch J.D., Zubizarreta, J, R., Stevens, J., Song, Z. (2025). Hospital Staffing and Patient Outcomes After Private Equity Acquisition. *Annals of Internal Medicine*, 178(11): 1529-1538

Kelly, C. J., et al. (2019). Key challenges for delivering clinical impact with AI. *BMC Medicine, 17*(1), 195

Knuth, D. E. (1997). *The art of computer programming* (3rd ed.). Addison Wesley

Kitchin, R. (2014). *The data revolution: Big data, open data, data infrastructures and their consequences*. Sage

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444

Madelena, Y Ng., Supriya, K., Blizinsky, K., & Hernandez-Boussard, T. (2023). The AI life cycle: a holistic approach to creating ethical AI for health decisions. Nat Med. 2022 Nov;28(11):2247–2249. doi: 10.1038/s41591-022-01993-y

Mesko, B., et al. (2017). Digital health is a cultural transformation of healthcare. *mHealth, 3*, 38

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 220–229). doi:10.1145/3287560

National Institute of Standards and Technology. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)*. U.S. Department of Commerce

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

Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information.* Harvard University Press

Pearl, J. (2009). *Causality: Models, reasoning and inference* (2nd ed.). Cambridge University Press

Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring Fairness in Machine Learning to Advance Health Equity. Ann Intern Med. 2018 Dec 18;169(12):866-872. doi: 10.7326/M18-1990

Russell, S., & Norvig, P. (2021). *Artificial intelligence: A modern approach* (4th ed.). Pearson

Topol, E. (2019). *Deep medicine: How artificial intelligence can make healthcare human again*. Basic Books

Vyas, D. A., Eisenstein, L. G., & Jones, D. S. (2020). Hidden in plain sight—Reconsidering the use of race correction in clinical algorithms. *New England Journal of Medicine, 383(9), 874–882*. https://doi.org/10.1056/NEJMms2004740

Williams, D. R., Lawrence, J. A., & Davis, B. A. (2019). Racism and health: Evidence and needed research. Annu Rev Public Health. 40:105–125. https://doi.org/10.1146/annurev-publhealth-040218-043750

World Health Organization. (2014). WHO handbook for guideline development (2nd ed.). World Health Organization

World Health Organization. (2021). *Ethics and governance of artificial intelligence for health.* World Health Organization

Xiang, Alexa Q. and Tohyama, Takeshi and Bank, Alexander Cole and Bui, Quang and Gorijavolu, Rahul and Henao, John Anderson Garcia and Jaiswal, Nikhil and Kelshiker, Akshay and Madapati, Kaushik and Ordóñez, Sebastian A. Cajás and Patel, Milit and Prakash, Nina and Celi, Leo Anthony, Distributed Open Justice Oversight (DOJO): A Community-Driven, Modality-Agnostic Platform for Adversarial Evaluation of Health AI (April 29, 2026). Available at SSRN: https://ssrn.com/abstract=6676818