The Physicians' Charter for Responsible AI

A Practical Guide to Developing, Testing, and Using AI Tools in Clinical Practice

physicianscharter.ai

organized by MD+CALC

Table of Contents

The Physicians' Charter for Responsible AI

A Practical Guide to Developing, Testing, and Using AI Tools in Clinical Practice

Executive Summary

physicianscharter.ai

organized by MD+CALC

01. About the Charter

The origin of our Physicians' Charter stems from a growing concern among the physician creators of MDCalc about the rapid pace of AI and how it will be implemented in healthcare. In the absence of an existing resource that provided practical, clear guidance using real-world clinical scenarios and authored by frontline physicians, we assembled a diverse group of experts to create one. This document is the collective effort of practicing physicians across numerous medical specialties. We all share enthusiasm for AI's potential in medicine, and are steadfast in our commitment to its ethical, fair, and patient-focused implementation.

Delivering care to patients every day provides us with a unique understanding of the intricacies of healthcare, a perspective we consider essential in guiding AI's integration into our field. There is a true urgency from physicians to set safe boundaries and demand high expectations from AI in the clinical environment.

We hope this charter offers a practical, understandable, and accessible framework to guide all stakeholders. As physician leaders in this era of AI evolution, we must always prioritize the values and the welfare of our patients above all. This charter is our pledge to ensure AI in medicine is effective, ethical, and fundamentally patient-centric.

Since the first physician saw the first patient, doctors have always searched for, developed, and adopted new tools to improve patient care. From the advent of the stethoscope in the early 19th century to the development of advanced medical imaging like the ultrasound during the 20th, physicians have integrated technology to diagnose diseases more accurately and improve patient outcomes. Similarly, clinical scores and algorithms have enabled physicians to make better, more informed decisions through research and evidence-based care. Physicians have endeavored to use these medical advancements with a focus centered on patient welfare and trust.

Today, we stand at the threshold of a new era. Artificial intelligence, which we will define as "software or systems that are capable of performing tasks by learning from and analyzing language and medical data" — is increasingly present in society and could even become an essential component of medicine.

As authors of this document, we wish to emphasize our enthusiasm and excitement for the promises AI brings. We are not creating this charter out of fear. Instead, our motivation stems from our passion and belief in AI's transformative potential. If developed and deployed responsibly, we believe AI can help to re-center the practice of medicine around the patient — not the computer screen — bringing physicians back to the bedside in a way reminiscent of how medicine was practiced centuries ago.

The promise of AI is significant: improved disease diagnosis, personalized treatment plans, and overall enhancements to patient care. However, with these opportunities come numerous challenges and ethical considerations. How do we ensure that AI systems offer transparency and respect patient privacy? How can we reduce algorithmic bias to prevent exacerbating health disparities? How do we maintain the irreplaceable humanity in medicine while leveraging the benefits of AI?

As we grapple with these and other numerous ethical questions, our charter seeks to provide a collective response. Our position is unequivocal: **patients must continue to be the central focus of medical care, and the sanctity of the patient-doctor relationship must be upheld**. AI is a tool designed to support and augment the capabilities of the healthcare professional - not replace them. AI tools can amplify our skills, acting as a co-pilot, enabling us to work more efficiently and smartly and ultimately allowing us to focus more on what truly matters: our patients.

In this document, we discuss our mission statement and vision for how AI tools — from generative chatbot models to machine learning and neural networks — should be used meaningfully, ethically, and to the benefit of the patients we serve. We base these recommendations on the ethical principles physicians have relied upon for centuries. We have also given great thought as to how these new technologies may require us to focus on several new principles: transparency, accountability, equity, and most importantly, human-centered care.

We then present our 10 Rules of the Road for integrating AI into medical practice. These practical guidelines are designed to be easily understandable by those who work in healthcare and include examples from clinical practice, drawing parallels with our past as we navigate the future.

As you explore this outline, we hope it catalyzes thoughtful discussions and informed actions around integrating AI into clinical practice. We believe in a future where AI, like the stethoscope or ultrasound, becomes an indispensable, supportive tool that enhances our ability to provide the best care for our patients, with the doctor-patient relationship always at its core.

02. Our Core Values

Our core values build on the four pillars of medical ethics: autonomy, beneficence, non-maleficence, and justice.

Human-Centered Care: The priority is always to serve the patient's needs and preferences, focused on their values. This includes respecting patient autonomy and involving patients in decision-making; it also includes focusing on outcomes, diagnoses, and treatments that are relevant and impactful to patients and their lives. AI should be used to enhance — not replace — the patient-doctor relationship and help physicians provide more personalized, effective, and efficient care.

Transparency: Clear and open communication about how AI tools function, how they were developed, and what data they use. This includes sharing how these tools impact clinical decisions and patient care.

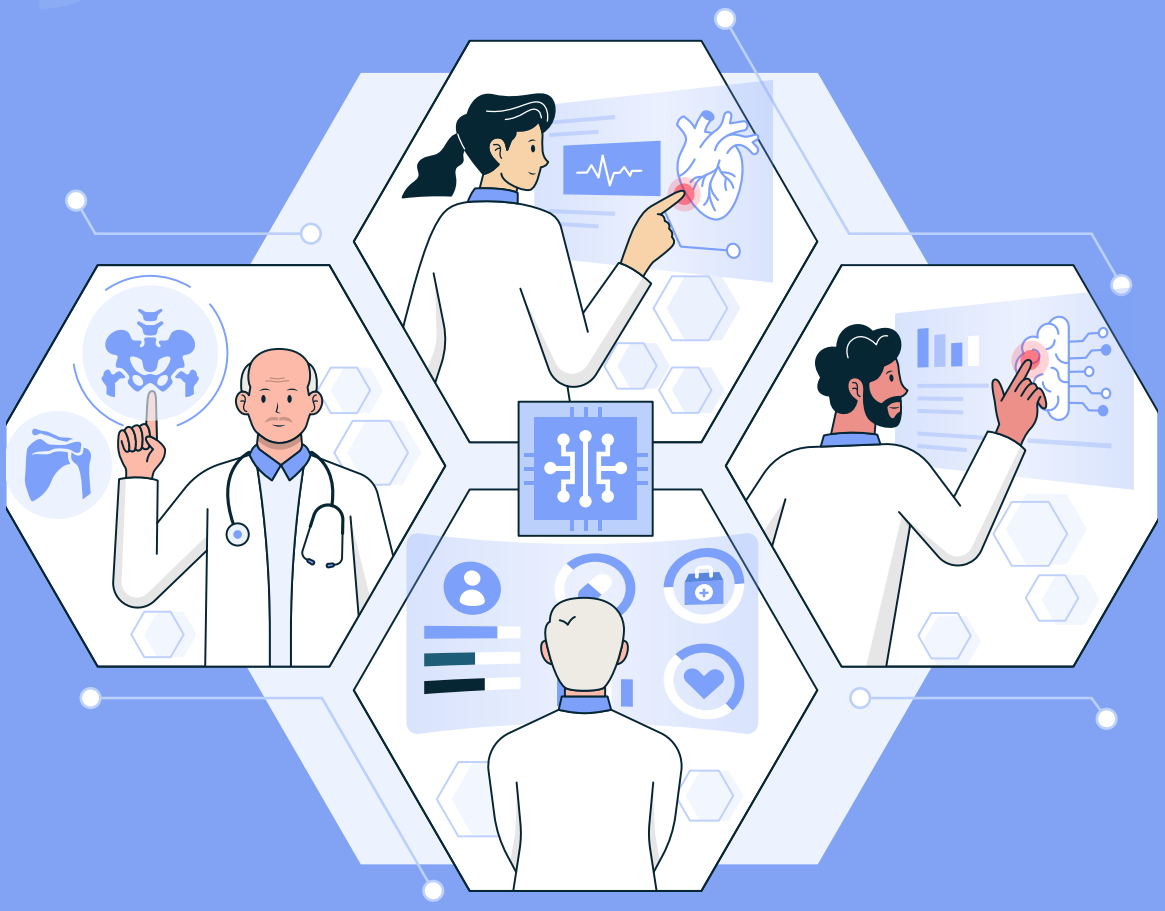
Privacy and Security: Safeguarding sensitive patient information by employing robust data protection measures and ensuring adherence to relevant laws and regulations.

Equity: Ensuring AI tools do not exacerbate health disparities but instead work to promote equitable care and outcomes. It means recognizing and acknowledging bias that exists today, prioritizing bias mitigation, demanding diverse data representation, and deploying AI equitably.

Collaboration: Engaging a variety of stakeholders in AI development and use, such as physicians, other healthcare professionals, data scientists, ethicists, and patients themselves. This collaboration facilitates a multidisciplinary approach and a more holistic view of patient care.

Accountability: Responsibility for AI's implementation and the outcomes it generates. This includes the need for regulatory oversight, accountability of systems not individuals, malpractice clarification and reform, adherence to privacy laws, and safety and error management. Additionally, we must consider the downstream effects of AI implementation: Are resources allocated differently? What are the unintended consequences and impact on patient care?

Continuous Learning and Improvement: Embracing a growth mindset, in which there is continuous monitoring of AI tools, ongoing validation, and an open environment for learning and adaptation. This includes both the AI systems and the healthcare professionals using them.



03. The 10 Rules of the Road for AI Implementation

Topic and Description	Examples in Clinical Practice
01  **Human-Centered Design and Engagement** Keep the patient-doctor relationship central, involve patients and doctors early in the development process and inform them about how AI is utilized in their care.	• In developing an AI tool for assessing depression, feedback from patients and psychiatrists is included in development from the outset to ensure the tool is both clinically useful and user-friendly. • During an AI-guided surgery, the surgeon discloses and explains to the patient the role of AI in assisting but not performing the procedure. • An AI diagnostic tool is used by a physician during the patient visit, so that the doctor can explain and discuss the tool with their medical expertise.
02  **Data Quality and Privacy** Prioritize high-quality, diverse, and geographically relevant data for training AI models. Respect patient privacy and foster responsible data interpretation.	• An AI tool for predicting disease progression should use diverse datasets representing different geographic and demographic cohorts, ensuring its efficacy across a broad patient population. • EHR data used to train AI models for outcome prediction should be de-identified and encrypted end-to-end to maintain patient privacy.

Topic and Description	Examples in Clinical Practice
03  **Ethics, Bias Mitigation, and Their Implications** Expect, monitor and mitigate biases in AI algorithms and consider potential ethical implications in AI deployment.	• In developing an AI tool for skin cancer detection, the model is trained on a diverse dataset representing various skin types to minimize bias. • When deploying an AI tool for prioritizing patient referrals, consider its impact on access to care to ensure it does not inadvertently favor or disadvantage certain patient groups.
04  **Trust: Transparency, Explainability, and Accountability** Encourage a “glass box” approach to AI, provide clear information about its workings, and establish a robust framework for trust and accountability.	• When using an AI model for predictive analytics, both patients and physicians are provided with clear, understandable explanations of how the model works, what data it uses in its analysis, and how it makes predictions (some AI models make this not possible). • An accountability framework is implemented so that in case of misdiagnosis by an AI tool, there are mechanisms for addressing the error and preventing recurrence through rapid feedback from clinician to AI developer.

Topic and Description	Examples in Clinical Practice
05  **Continuous Validation, Feedback, and Improvement** Ensure that AI tools are evaluated on and iterate from formal, objective evaluations of their utility as well as in everyday use; models require continuous review. Encourage feedback and provide clear paths for users to share their experiences and insights, keeping the tools effective, safe, and up-to-date.	• An AI tool for diagnosing diabetic retinopathy is repeatedly and regularly validated on diverse, independent datasets. Its performance is closely monitored over time using standardized benchmarks. Any erroneous suggestions from this AI tool are directly reported by physicians via a dedicated feedback system, leading to the tool's refinement and improvement. • An AI system built for heart disease diagnosis is not just validated once, but regularly checked against standard performance measures and monitored for drift in accuracy. Doctors using the system can report any inaccuracies they find, helping make the system better over time.
06  **Collaborative Approach and Workflow Integration** Promote a collaborative, compensated, multidisciplinary approach to AI development, focusing on AI tools that integrate seamlessly into healthcare workflows.	• In developing an AI tool for radiology, radiologists, data scientists, ethicists, and patients are all involved, with compensation structures in place for the time required to review these models. • An AI tool for analyzing CT scans is designed to integrate directly into a hospital's existing imaging and EHR systems, providing insights within the existing workflow.

Topic and Description	Examples in Clinical Practice
07  **Regulatory Compliance and Safety** Adhere to regulatory guidelines for AI development and implementation. Implement robust safety measures to protect patient safety.	• AI-based diagnostic tools are developed in accordance with guidelines and regulations, ensuring their safety and efficacy. • In AI-assisted surgery, backup safety measures are considered and ready for use to prevent potential harm from AI-induced errors.
08  **Education and Support** Provide comprehensive education and training to healthcare providers about AI, and support them in their roles as primary interpreters of AI outputs.	• In a hospital deploying an AI tool for radiology interpretation, a comprehensive training program is provided, offering radiologists extensive knowledge about the tool, its use cases, and how to interpret and verify its outputs. • A healthcare organization provides an ongoing support program for clinicians, providing regular updates, resources, and direct communication lines with the AI development team for queries and feedback.

Topic and Description	Examples in Clinical Practice
09  **Patient-Centered Outcomes and Value in Healthcare** Develop clinically meaningful AI tools that enhance healthcare value, reduce overdiagnosis and overtreatment, and provide better outcomes at the same or lower cost.	• An AI tool for lung cancer screening is trained to accurately differentiate between benign and malignant nodules, thus reducing unnecessary invasive procedures and patient anxiety. • An AI system for detecting pulmonary emboli takes into consideration that some emboli may be clinically insignificant (or even false positives). It includes the risk of anticoagulation and PE treatment into its model, recognizing that patient-important outcomes are the overall goal, not just detection of clot.
10  **Understand the Limits of AI** Recognize that while AI can augment and improve healthcare delivery, it is not a panacea and cannot solve every problem in our complex, fragmented healthcare system. We must understand its limitations, understand when human intervention is needed, and find a balance between technological assistance and human action to provide optimal care to patients.	• AI models can assist in predicting disease progression based on extensive data sets, but these predictions are purely statistical and do not account for individual patient responses and differences. Clinicians must interpret these predictions while considering their personal understanding of the patient's condition and unique circumstances. • An AI algorithm may be capable of sorting and prioritizing patient referrals based on their medical data, but it cannot wholly substitute the human touch in empathizing with patient fears and anxieties. Clinicians are needed to communicate comfort and provide the caring human interaction that patients often require.

04. Rules of the Road: In Depth

Rule 1: Human-Centered Design and Engagement

William Collins, MD

Preserving and supporting the patient-doctor relationship must be a central tenet of design, testing, and implementation of AI systems for medicine. At its core, the medical profession exists as a service from one human to another. Unfortunately, modern medicine often emphasizes efficiency and profit and AI has the potential to both push medicine further away from the primacy of healing. Yet it also has the potential to help restore this essential dogma. We actively seek the latter. AI systems should be designed to augment the essential human interaction in medicine rather than replace it. Providers and patients should be engaged participants in this process.

Understanding Human-Centered Design and Engagement

Human-centered design approaches problem solving with the human user as the primary focus, seeking to optimize design around the needs of the user. In the context of AI applications for medicine, human-centered design places both the patient and the patient-provider relationship at the center of the design.

Human-centered engagement refers to the importance of patients and providers being directly involved in the design of AI systems. Early feedback from these primary users is essential to the success of any system in its ability to serve its users. AI should never be designed to replace the patient-provider relationship – but can absolutely enhance and facilitate it.

Why is Human-Centered Design and Engagement Essential?

Medicine is and must remain a fundamentally human endeavor. While some may suggest that AI systems can express human empathy in written responses, there is a large gulf between a written message and an in-person interaction. Centering the design of healthcare AI around patients and their providers means focusing on solutions that increase the time providers can spend directly with patients. This could come in the form of early triage systems to direct patients to the correct provider and applications to ease laborious documentation, generate accessible visit summaries, and help with responsive messaging to patients.

To best center the human patient in AI design, designers must also engage early with patients and providers to ensure systems are serving the correct needs and are highly functional. Care should also be taken not to bias toward more affluent groups in engagement. The goal of AI in healthcare should be to lift and improve the human condition across spectrums.

Ensuring human-centered engagement requires clear disclosure to both patients and providers when an AI system is used. This should include a comprehensive understanding of how and why an AI tool is being utilized. The primary aim of these systems should not be purely efficiency or profitability. Rather, the design and deployment of AI should be oriented around human interaction and the fundamental ethos of care that underpins medicine. This approach is crucial to prevent other objectives from dominating the development and application of AI.

Human-Centered Design and Engagement in Action

Let's consider some examples to illustrate how human-centered design and engagement is best utilized in AI solutions for healthcare:

AI in Secure Patient Messaging: AI systems designed for patient messaging platforms can aid in healthcare democratization and increase patient response speed while reducing clinician burnout. Systems like these could even allow the patient to ask near-infinite additional questions, and AI systems could provide personalized, relevant responses at any time of day or night — an unrealistic expectation for human physicians.

AI in Undiagnosed Disease: An AI diagnostic tool is used in the context of a patient's evaluation by a physician so that the physician can interpret the tool using their medical expertise; the physician understands the limitations of this tool, and the tool is designed so that its recommendations focus on "patient-important outcomes" — outcomes that are important to humans wanting to live full and healthy lives, not surrogate markers.

AI in Mental Health: In developing an AI tool for assessing depression, feedback from patients and psychiatrists is included in its initial development to ensure the tool is both clinically useful and user-friendly.

Conclusion

To create the most effective AI systems for medicine, the focus must be on maintaining (and in many ways restoring) the primacy of the patient and provider relationship. These human users should be involved and informed throughout all stages of development to optimize AI solutions for their care.

Rule 2: Data Quality and Privacy

Dustin Cotliar, MD, MPH and Anthony Cardillo, MD

"Garbage in, garbage out" is a principle of all data-driven models, and in AI, it is no different: AI output quality depends entirely on data input quality. Much like medical students' knowledge is shaped by their education and experiences, AI models learn from the data they are given. If trained on diverse, high-quality data, the AI — like the experienced physician — can make accurate predictions across varying scenarios and patient demographics. Conversely, if trained on low-quality, disorganized, or homogenous data, the AI may yield inaccurate or biased results, much like a doctor with limited exposure or inadequate training.

While data quality is key, so is data privacy; we fully recognize that AI tools are only possible by compiling large sums of high-quality data, but each piece of that data is derived from an actual person. Critical to responsible and ethical AI development is that this compiled data must be anonymized/de-identified. If patients cannot trust that their data is secure, they may hesitate to share critical health information, leading to substandard care and medical errors. At the very least, AI tools that access one's medical data must not have any negative consequences to any individual. In addition, we must also recognize that we likely all have individual biometric fingerprints (retinal image, heart rate patterns, molecular makeup) that could identify individuals and that information must be protected as well.

The Challenges of Data Quality and Data Privacy

Gathering high-quality, diverse, and geographically relevant datasets is a multifaceted challenge. Furthermore, stringent data privacy laws protecting sensitive health data may make it more challenging to access high-quality data. Poor interoperability between different healthcare systems and the over- or under-representation of certain demographics present struggles as well. Significant cost and time are required to collect, clean, and curate these datasets — especially on a global scale.

Ethical and safe implementation of AI models in healthcare will require overcoming these challenges in the most transparent way possible. A multi-pronged approach includes fostering more collaborative data collection efforts, enhancing collection and cleaning techniques, and promoting interoperability through universal standards and regulations. Lastly, fostering transparency about potential biases and limitations inherent in these datasets is paramount. Open dialogue about these challenges not only bolsters confidence in AI tools among clinicians and patients but also promotes proactive discussion around mitigating any adverse effects on patient care.

The risk of violating individual privacy and eroding trust in healthcare AI is substantial if sensitive patient data is not stringently protected. The robust data sets powering AI models will necessitate unprecedented privacy and security safeguards not previously seen in healthcare. As we enhance AI models with diverse, high-quality data, we must bolster our approaches to privacy control and responsible data interpretation.

AI Tool	Impact of Low Quality, Homogenous Data Sets
Diabetes Risk Prediction Model designed to predict the risk of developing diabetes based on individual health metrics (e.g., age, BMI, blood pressure).	**Inaccurate risk estimations**. Models trained primarily on data from older, obese patients, might overestimate the risk for younger, healthier individuals and underestimate it for different demographic groups. This could lead to inappropriate interventions or missed preventative measures.
Radiology AI Model designed to read and interpret radiologic images (like CT scans or X-rays) to detect conditions like lung nodules or brain tumors.	**Inaccurate diagnoses**. If trained on lower-quality images or primarily images from one patient group, the model may fail to accurately identify conditions in diverse populations. For instance, it might miss certain lung nodules more prevalent in a specific ethnicity or misinterpret normal variations as abnormal findings, which can lead to unnecessary interventions and patient distress.
AI Model for Mental Health Assessment used to evaluate patients' mental health status based on factors like speech patterns, tone of voice, text inputs, and facial expressions.	**Misinterpretation of cultural norms**. If the model is trained on data from a limited cultural or geographic population, it may misinterpret cultural idiosyncrasies as signs of mental health issues. This could lead to over-diagnosis or inappropriate treatment recommendations which could worsen biases and healthcare disparities.

Robust Privacy Controls and Responsible Data Interpretation

AI Tool	Privacy Impact	Data Interpretation Impact
AI Model for Genomic Medicine that uses genomic data to predict disease risk.	Patient genomic data could be misused for discriminatory practices (e.g., by employers or insurers) or unauthorized research.	Model outputs could lead to alarming predictions about disease risk, causing unnecessary patient distress and possibly unnecessary medical interventions.
AI Model for EHR Analysis that extracts information from electronic health records (EHRs) to predict health outcomes.	Improperly de-identified patient data could lead to identity theft, leaking of personal details in the model, or misuse of personal health information.	Clinical predictions that are not properly contextualized within the patient's overall health status might result in inaccurate recommendations and medical errors.

Future Implications for Data Quality and Privacy in Healthcare AI

The swift integration of AI in healthcare offers promising opportunities while presenting certain challenges. It's crucial to recognize the hurdles we face in guaranteeing data quality and patient privacy, as both are vital to the success of AI implementations. Without careful attention to these facets, we risk inaccurate diagnoses, data breaches, and broader cybersecurity threats.

Data quality is an absolute necessity for AI performance — and lots of it. Incomplete, outdated, or biased data compromise the precision of AI models, possibly leading to over- or under-diagnosis or treatment.

Equally important is adequate data privacy. Aggregated data—although immensely valuable for AI—must be protected rigorously to avoid unintended disclosure of sensitive patient information. This is especially important with the rise of numerous "biometric fingerprints" that are unique to each individual person. Robust privacy controls are critical, requiring the use of de-identification, anonymization, and encryption techniques. Still, even with these methods in place, the risk of information leakage persists.

While end-to-end encryption is a widely available technique and should be implemented as a baseline, more advanced privacy techniques such as homomorphic encryption, differential privacy, and secure multi-party computation can offer additional safeguards. Homomorphic encryption facilitates computation directly on encrypted data with no intermediate decryption. Differential privacy can be used to "hide" individuals in a dataset by applying carefully constructed noise during the training process. Lastly, secure multi-party computation allows multiple parties to jointly train machine learning models without providing their private datasets. These newer techniques are especially effective for securing data in "federated learning," a popular training paradigm in which a single AI model can be trained across multiple institutions and their local data.

Conclusion

As with many of our Rules of the Road, inherent ethical conflicts must be navigated carefully when implementing AI models into patient care. Because stringent privacy protections and higher data quality standards might slow the rate of change and reduce model accuracy, there could be a tendency to forgo essential precautions in these areas. For these reasons, we must always return to our critical values and principles to help guide the way: the patient is at the center of their healthcare needs, and we must tirelessly focus on their wellness and privacy while delivering their care in an AI-enhanced practice of medicine.

Rule 3: Ethics, Bias Mitigation, and Their Implications

Sarah Gebauer, MD

All patients deserve excellent care regardless of their personal attributes, and it is our responsibility as physicians to advocate for AI prioritizing patient welfare. In 2021, the World Health Organization (WHO) described key ethical principles (Figure 1) in addition to considerations for regulation, governance, and public engagement.

Table 1: Key ethical principles for use of AI in healthcare

1. *Protect autonomy*
2. *Promote human well-being, safety, and the public good*
3. *Ensure inclusiveness and equity*
4. *Ensure transparency, explainability, and intelligibility*
5. *Foster responsibility and accountability*
6. *Promote AI that is responsive and sustainable*

The first three principles are foundational to medical care and are familiar to most physicians. The last three principles relate more specifically to AI and can be applied to a broad range of actions and concerns. Bias is particularly insidious, as it can be challenging to detect and measure, and can affect each of the ethical principles. Given that the consequences of bias can directly impact each of these ethical principles, there is a strong ethical imperative to actively mitigate bias in AI.

Even in its limited debut into healthcare, there have already been clear examples of AI bias:

Lack of Representation in Datasets: A 2022 Lancet study reviewing all publicly available dermatology image databases found that skin lesions in darker-skinned populations were markedly underrepresented, possibly leading to poorer AI performance.

Racial bias in Resource Allocation: Science's 2019 paper found that Black patients were less likely to qualify for additional support programs despite being sicker than White patients. Researchers determined that the algorithm relied on healthcare costs as a proxy, and Black patients had lower healthcare costs with the same amount of illness.

Gender Bias in Clinical Risk Assessment: In a Nature study, researchers trained a deep learning kidney injury prediction algorithm on data from the Veterans Administration (VA) with data that was 94% male. The resulting risk assessment performed significantly worse on female patients.

Types and Instances of Bias
AI bias can be systemic or data-related.

These biases occur in addition to standard human bias, which is unlikely to be completely resolved with a technological approach. They require ongoing vigilance and understanding by physicians and the healthcare team to understand their impact on patients and healthcare.

Systemic bias occurs when the AI model reflects the biases of society, healthcare, and physicians and perpetuates the current system's injustices.

Data bias occurs when groups of people are not well-represented in a model's training data, which can result in inappropriate suggestions or actions.

Human bias includes cognitive and confirmation bias which may affect an AI model's function. Cognitive biases can unintentionally influence how data is interpreted, potentially skewing the model's outputs. Confirmation bias, on the other hand, may lead to over-reliance on the model's predictions that confirm preconceived notions while neglecting contrary indications. These biases can lead to misinterpretations, inaccurate diagnoses, or inappropriate treatment recommendations.

Type of Bias	Clinical Examples
Systemic Bias	
Historical	Clinical decision-making systems are influenced by outdated historical practices, such as differential treatment plans based on race or gender due to historical biases.
Societal	Societal biases such as the stigma surrounding mental health can influence AI systems, possibly resulting in underdiagnosis or undertreatment of mental health conditions.
Institutional	An AI system might prioritize patients who have private insurance over those who are uninsured or have public insurance due to institutional biases in care delivery.
Data-Related Bias	
Model selection	Choosing a model that was primarily trained on urban patients might not perform as well for rural patients, leading to inaccurate predictions.
Survivorship	An AI developed to predict cancer survival based on data from patients who survived may not accurately predict outcomes for patients with more aggressive forms of cancer.
Data dredging	An AI model meant to predict the risk of developing diabetes sifts through large volumes of health data without a defined hypothesis or strategy. It might start identifying patterns linked to unrelated factors such as hair color or favorite food. These correlations may simply be coincidences within the dataset, but the AI could incorrectly consider them important, leading to unreliable diabetes risk predictions.
Representation	An AI model used to make recommendations for preventive healthcare screenings (such as mammograms or prostate exams) might be based largely on gender data tied to an individual's sex assigned at birth. This might lead to incorrect screening recommendations for transgender individuals, potentially missing critical preventive care opportunities.

Type of Bias	Clinical Examples
Human Bias	
Availability heuristic	If a clinician has recently seen many cases of a rare disease, they might overestimate its prevalence, and this could bias an AI system trained on their diagnostic decisions.
Confirmation bias	An AI system trained to predict patient outcomes based on clinician's notes might inherit clinicians' biases if they tend to interpret information in a way that confirms their preconceptions.
Implicit bias	An AI model trained on data from clinicians who, unknowingly, provide less aggressive treatment to certain groups (such as elderly patients or racial/ethnic minorities) might mirror these discriminatory practices.

Within each of these biases, issues arise related to **datasets, processes, and monitoring.**

- Datasets determine who is counted and relate to issues of sampling bias, underrepresentation of marginalized groups, and human decisions about data availability.
- Processes determine how to optimize the model (i.e., in favor of the minority or majority), how the variables collected impact the model, and the humans determining the optimal outcomes.
- Monitoring determines how the model is functioning optimally and relates to changing the model with changing societal norms, survivorship bias, and confirmation bias.

This figure from the National Institute of Standards and Technology (NIST), demonstrates the ways biases contribute to harm:

Figure 3: How Biases Contribute to Harms, from NIST's Towards a Standard for Identifying and Managing Bias in Artificial Intelligence

	Systemic Biases	Statistical and Computational Biases	Human Biases
Datasets Who is counted, and who is not counted?	Issues with latent variables Underrepresentation of marginalized groups	Sampling and selection bias Using proxy variables because they are easier to measure Automation bias	Observational bias (streetlight effect) Availability bias (anchoring) McNamara fallacy
Processes and Human Factors What is important?	Automation of inequalities Underrepresentation in determining utility function Processes that favor the majority/minority Cultural bias in the objective function (best for individuals vs best for the group)	Likert scale (categorical to ordinal to cardinal) Nonlinear vs linear Ecological fallacy Minimizing the L1 vs. L2 norm General difficulty in quantifying contextual phenomena	Groupthink leads to narrow choices Rashomon effect leads to subjective advocacy Difficulty in quantifying objectives may lead to McNamara fallacy
TEVV How do we know what is right?	Reinforcement of inequalities (groups are impacted more with higher use of AI) Predictive policing more negatively impacted Widespread adoption of ridesharing/self-driving cars/ etc. may change policies that impact population based on use	Lack of adequate cross-validation Survivorship bias Difficulty with fairness	Confirmation bias Automation bias

Figure: How Biases Contribute to Harms, from NIST's Towards a Standard for Identifying and Managing Bias in Artificial Intelligence

Awareness of bias in AI models is crucial, and ideally a multidisciplinary group of stakeholders is involved in the entire AI development lifecycle including concept generation, model design, and monitoring. Specific consensus guidelines are evolving. Even when a bias has already affected an AI model, healthcare professionals have a role in addressing its impact on patient outcomes.

Some concrete steps to mitigate data bias include:

Inclusive Data Collection: Ensure that data collection processes are designed to capture a diverse range of patient attributes, including race, ethnicity, gender, socioeconomic status, and geographic location.

Multi-Disciplinary Team Review: Teams planning on implementing AI technologies into healthcare should have broad, diverse representation to ensure that many perspectives are represented.

Rigorous Data Cleaning and Preprocessing: Prioritize rigorous data cleaning and preprocessing techniques to remove or mitigate biases present in the data before training AI algorithms. This includes using statistical methods to balance group representation in datasets.

Transparent Algorithm Design: Transparent algorithmic design enables healthcare professionals to understand how the AI system arrives at its predictions or recommendations. By providing explanations for the decisions made, clinicians can validate the outputs, detect potential biases, and build trust in the AI system. The transparent design also allows for external audits and scrutiny, enabling experts to identify and rectify biases more effectively.

Regular Algorithmic Updating and Monitoring: Healthcare AI algorithms should be regularly updated and monitored to ensure they are continuously learning from new data and adapting to changing contexts. Ongoing monitoring helps identify any emerging bias in real-world usage, and automated dashboards can help scan for common types of biases in data. Regular updates also provide an opportunity to address biases identified through evaluations and audits.

Conclusion and Steps to Mitigate Systemic and Human Bias

Ultimately, creating a more fair and just society and healthcare system is the best way to mitigate bias.

Acknowledging these biases, systemic, data-related or human, is the first step towards mitigating them. It is not if there is bias, but "Where is it, and how can we do our best to address it safely and ethically?"

Inclusive data collection, diverse multidisciplinary team review, rigorous data cleaning and preprocessing, transparent algorithm design, and regular algorithmic updating and monitoring are key strategies for bias mitigation. These efforts should be reinforced by diverse, multidisciplinary team reviews that actively involve all stakeholders, including patients. This participatory approach helps to ensure that the AI systems developed are truly reflective of and beneficial to the diverse populations they serve.

At the same time, we must acknowledge the significant role of policy and regulation in addressing bias. As AI evolves and becomes more integrated into healthcare, so too must our legal and regulatory frameworks adapt to ensure fair and equitable AI use. (See Rule 7, Regulatory Compliance and Safety.)

Mitigating bias in AI is a continual process. As our knowledge expands, as AI systems learn and evolve, and as societal norms change, our strategies for identifying and addressing bias must also evolve.

Rule 4: Trust: Transparency, Explainability, and Accountability: Unveiling the Inner Workings of AI

Graham Walker, MD

Transparency, explainability, and accountability are vital for trust in healthcare AI systems, and without trust, AI models will simply go ignored, unused, and abandoned by patients and providers alike. When physicians and patients understand how these systems operate and there are clear mechanisms for handling mistakes, trust is fostered, enhancing the adoption and effective use of AI tools in healthcare.

Understanding Transparency, Explainability, and Accountability in AI

Transparency in AI refers to information about the data used for training models, the decision-making algorithms, and validation methods. On the other hand, explainability involves the AI model's ability to provide comprehensible justifications for its predictions or decisions. Given the complexity of some AI models, like deep learning or neural networks, achieving total transparency and explainability can be challenging, but it remains a vital objective, especially in high-stakes fields like medicine. Accountability, meanwhile, involves assigning responsibility when an AI model leads to an error, especially one that could harm a patient. This means defining roles for different stakeholders, including AI developers, healthcare providers, and organizations implementing AI tools.

Why Transparency, Explainability, and Accountability Are Essential

Medicine is a mixture of art and science, and when evidence is lacking in medical practice, physicians rely heavily on their understanding of the workings of the human body and disease: anatomy, physiology, pathophysiology, and pharmacology. Because of this, a physician can explain their decision-making process. However, many AI models may lack this ability, acting as "black boxes" due to their intricate algorithms. Such opaqueness can lead to skepticism among physicians, potentially impacting the adoption of AI in healthcare.

In these cases, consistent performance and accurate decision-making becomes paramount for building trust. As a parallel, when new medical technologies such as ultrasound or CT scans were first introduced, there was a high demand for explainability until clinicians became familiar with their capabilities. Over time, this demand for explainability waned as trust in the technologies grew. (Several new techniques may address this challenge, like Explainable AI/XAI and models that explicitly train "Explanation Algorithms" alongside black box algorithms.)

Mistakes are inevitable, and the way they are handled is crucial. An accountable AI system needs mechanisms for identifying, rectifying, and learning from errors. Developers and healthcare providers share this responsibility: developers ensure the AI model's proper functioning, while healthcare providers must use the tool correctly, interpret its outputs accurately, and consider its recommendations alongside their clinical judgment. Additionally, regulatory compliance, such as that required by the FDA, necessitates AI tools to have defined accountability structures and to demonstrate processes in place to handle errors compliantly. Finally, we cannot simply maintain the current system wherein the physician is singly responsible for challenges that arise during the course of medical practice, particularly now that evidence demonstrates that providers tend to anchor on diagnoses made by AI in some processes.

Examples of Transparency, Explainability, and Accountability in Action

AI in Radiology: Consider an AI tool used to detect lung nodules in CT scans. Its effectiveness depends on its ability to highlight anomalies and explain its reasoning, such as the size, shape, or density. Suppose the AI tool were to make a wrong prediction: accountability would involve flagging the error. In that case, developers investigate the cause, and healthcare organizations implement checks to prevent similar errors, and contact patients impacted, thus underlining the shared responsibility and mechanisms for error correction.

AI in Predictive Healthcare: An AI tool correctly predicts the progression of neurodegenerative diseases like Alzheimer's. While the complex deep learning model utilized might make complete explainability challenging, it could provide a degree of transparency by sharing key influential factors, such as genetic markers or patterns identified in brain imaging data. However if physicians are unable to understand or trust the model's output due to its complexity, they might be reluctant to use it, potentially limiting its utility in practice. This underlines the importance of a balance between sophisticated modeling and sufficient transparency and explainability to maintain user trust and acceptance.

AI in Oncology: AI models predicting a patient's chemotherapy response based on their tumor's genetic profile should clearly state the genomic markers it is using and their significance. This promotes transparency and explainability and aids oncologists in personalizing treatment plans.

Conclusion

To realize the full potential of AI in medical care, a steadfast, principled approach is essential. Transparency and explainability ensure that AI models can justify their decisions in understandable terms, building trust among healthcare providers and patients and fostering an environment conducive to the responsible use of AI in medicine. With the evolution of trust in AI systems, the explicit need for explainability may diminish but should never be completely discounted. Changes in AI model performance, unexpected outcomes, or new applications may require a renewed focus on explainability. The principle of accountability holds paramount importance as well. It not only defines who is responsible when AI tools make a mistake but also ensures that AI systems are designed and used responsibly. As we further integrate AI into healthcare, this principle becomes increasingly critical. When errors occur, they need to be managed in a manner that prioritizes patient safety and promotes continuous learning and improvement.

These three principles form a powerful triad. They foster trust, enable effective communication, assure responsibility and growth within AI-enabled healthcare, and provide a solid ethical foundation.

Rule 5: Continuous Validation, Monitoring, and Improvement

Carly Eckert, MD, MPH and S. Morgan Jeffries, MD

Okay, so you've finished building (or purchasing) your ML model, you've figured out how to integrate it into the workflow, and you've successfully deployed it. Congratulations! It's time for you to kick back and enjoy the fruits of your labor. Sorry, correction: most of the work is still ahead of you. You're now entering the long tail of the ML lifecycle: continuous monitoring. This is unlike many scores and algorithms we're used to — once published, the researchers often move onto a new question. This is not the case in most areas of machine learning and AI. Let's find out why.

Data: The Source of All Your Problems

The journey of your data is as important as the data itself: where it originated, when it was collected, the transformations it underwent to train your model—these elements remain crucial even post-deployment. Now that your model is in active use and your data pipeline is potentially processing new data in near real time, it's paramount to ensure the quality and consistency of the data you're scoring.

Are all the features in your model available at the time of scoring? It's easy to overlook the time lag in features like diagnosis codes when engrossed in model building. Is the origin of your training data congruent with the data you're now scoring? For instance, if you trained your model on data from an ambulatory care clinic, is it performing adequately in an emergency department setting? Or was your training data collected during a specific event such as the height of COVID? How can you ensure that the data you are scoring has undergone the same treatment as your training and validation data? Healthcare data is notoriously sparse; does the sparsity in your scoring data align with your training data?

Shifts and Drifts

Dataset shift, also known as domain shift, occurs when the distribution of data seen by the deployed model differs from what it encountered in training. For example, say you open a new pediatric emergency department, causing your volume of pediatric patients to grow overnight; the data available to your system has now drastically shifted.

When dataset shift develops gradually, it can result in model drift, which is the deterioration of a model's performance over time. Model drift comes in two main varieties: data drift and concept drift.

- Data drift is when the distribution of input data changes over time.
- Concept drift is when the relationship between the inputs and the target (the thing being predicted) changes.

The Labels: Easier Said than Done

Many of the AI applications that you will likely be involved in use machine learning (and supervised learning to be specific). With supervised learning, there is a labeled outcome (such as 'sepsis') that the model is learning to predict. But how do these labels get assigned? While sometimes these labels are self-explanatory (a discharge disposition of 'to hospice', for example), others are much more complex. How might a label diagnosing diabetes be developed? This could include diagnostic codes, lab results, or medications — with numerous caveats. Many labels are far from straightforward. It's crucial to remain consistent in label development from training, to validation, to labels used in the wild.

Another important check is prevalence of disease or condition, usually the percentage of patients (or encounters) with the particular condition. Consider the case where 5% of your training data has a wound infection following abdominal surgery. Upon evaluation of your first batch of test data, the prevalence increases to 12%! Such a discrepancy should prompt an evaluation of both how the training label was defined as well as the characteristics of the studied populations. There is potentially either a problem with your model or an error in your cohort development.

What To Monitor

Tracking model performance metrics is the easy part. Ideally you have a nice model dashboard, complete with an array of model metrics (see below). However, metrics alone are insufficient. No model is perfect, so how are false positives and false negatives handled in your clinical environment? And how are they being tracked? One of the advantages of machine learning is that models can improve over time, but even this does not happen automatically. The establishment of feedback loops to capture model errors is imperative so that future model iterations can improve.

The Data: It's essential to monitor your input data and target variable(s) to identify significant shifts over time. If such shifts occur, your model may not perform as well on the new data distribution. Track data provenance (the origin and history of your data) and watch for red flags. For example, using data from elective surgeries collected from May to July 2020 as your training data could be problematic due to the potential impact of the COVID-19 pandemic on such procedures. Furthermore, as the data that we track grows over time, stronger trends toward certain morbidities, demographics, or outcomes may emerge that are different from the ones we initially intended, which may again shift our compass in the model's target audience

Performance Metrics: Continuously monitor the performance metrics of your model on new data to check for degradation in performance over time. Some common metrics you might use:

- Sensitivity/Specificity, measures of a test's ability to detect disease or negate detection of non-disease
- Positive and Negative Predictive Values, probabilities that patients have or don't have a disease, based on a positive or negative screening test
- Accuracy, a measure of the degree that a model's predictions are correct
- F1 Score, a measure of a model's accuracy in a dataset (especially when true positive and true negative tests are uneven)
- AUC (Area Under the Curve), a measure of model fit relative to the gold standard or "perfect" test

Finally, be wary of any report that provides one single metric to evaluate model performance. Remember our example earlier of training data with a 5% prevalence of wound infection following abdominal surgery? A dataset with a 5% positive class (and 95% negative class) is highly imbalanced — as are most datasets in healthcare. The condition or event we are often interested in is usually uncommon. A model trained on the wound infection data would be 95% accurate by simply always predicting 'no infection,' but this model would have a sensitivity (also called recall) of 0%.

Human Review: Regular manual reviews of a sample of predictions from your model in production by human experts can provide unique insights into the model's performance. These experts can identify changing patterns of errors or validate the reasonability of predictions. Speaking of humans...

Governance and the Role of Clinicians

Clinicians should play a central role in multidisciplinary teams monitoring AI tools. Every AI model or tool used in clinical care should have a clinician "owner" and predefined metrics for accuracy and defining "success." Clinicians can also help prevent patient harm when "things go wrong" and facilitate the model's continuous improvement. If a problem is identified with a model, the clinician can help review how to prevent patient harm, how to review errors or problems with the model (and determining how or why it failed), and how to improve the model for the future (or decide to remove it from production entirely for review).

What To Do When There's a Problem

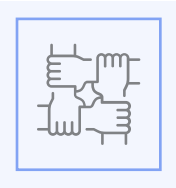
Model Re-training: Periodic retraining of your model on the newest samples and comparing the model metrics to your existing production model helps identify drift. Significant performance differences indicate a need to update your model.

Adapting: Along with re-training and retirement, adapting your model to fit new data patterns can also be an option. This could involve adjusting the parameters of your model or introducing new features that better capture the current data patterns.

Retirement: Despite your best efforts, some models will consistently underperform, others will cause workflow issues, and still others will simply be ignored by users. In each of these cases, retirement can be an easy win, both alleviating problems and reducing maintenance overhead.

Conclusion

The journey of an ML model in healthcare doesn't end at deployment. Continual validation, monitoring, and improvement are crucial to ensure the model's efficacy and reliability. By carefully watching data sources, model performance, and incorporating human review, we can navigate the long tail of the ML lifecycle and ensure our AI tools remain robust and beneficial to patient care.



Rule 6: Collaborative Approach and Workflow Integration

Matthew Sakumoto, MD

In AI development, the necessity of collaboration and seamless workflow integration cannot be overstated. While this collaboration might slow down the development process, it is critical for successful adoption and trust by users, including doctors and patients. This chapter emphasizes a multidisciplinary approach and AI tools' seamless integration into existing workflows. By doing so, we can foster collaboration and enhance existing healthcare practices.

The Need for a Multidisciplinary Approach

A successful AI integration demands diverse stakeholders' involvement—physicians, AI experts, ethicists, legal experts, patients, advocacy groups, administrators, and policymakers—each contributing their unique expertise. Physicians, with their domain knowledge and clinical insights, are essential for grounding AI solutions in real-world medical practice. AI experts and data scientists contribute technical knowledge and analytical skills to develop robust algorithms and models. Ethicists and legal experts ensure responsible AI practices by incorporating ethical frameworks and considering legal implications. Patients and advocacy groups provide critical perspectives, representing the voices and interests of those receiving healthcare. Administrators and policymakers shape policies and strategies, influencing the direction of AI implementation in healthcare systems.

A key part of this collaboration is the back-and-forth between clinicians and software developers/engineers. Clinicians bring critical information about patient care, while developers provide insights into technical possibilities and limitations. This exchange fosters mutual understanding and enables the creation of AI tools that are both technically sound and clinically relevant.

The benefits of a multidisciplinary approach are far-reaching. By involving diverse stakeholders at the outset, we can comprehensively consider medical, ethical, legal, patient-centric, and systemic aspects. This holistic evaluation promotes a balanced and responsible approach to AI development. Furthermore, collaboration enhances transparency and accountability, as stakeholders contribute their expertise and perspectives, fostering open dialogue and responsible decision-making.

Collaboration in AI Development

Establishing collaborative frameworks and partnerships is vital for AI development. Institutions can foster collaborations, uniting researchers, clinicians, and AI experts. These collaborations enable the sharing of knowledge, expertise, and resources, leading to advancements in AI applications in healthcare.

Encouraging and Incentivizing Collaboration

Recognizing and fairly compensating stakeholders encourages participation, ensuring their insights and expertise enhance AI development. This recognition encourages active participation and commitment from stakeholders. Fair remuneration ensures that the time, effort, and expertise invested by collaborators are duly acknowledged and compensated, further incentivizing their involvement in AI development. Collaborative approaches also increase adoption and acceptance among physicians and stakeholders, as they are actively engaged in the development process and can provide valuable insights and feedback.

Ethical considerations should also guide compensation models for collaborative AI projects. To avoid conflicts of interest, transparent processes must be established to ensure that financial arrangements do not compromise the integrity and objectivity of the development process. Additionally, alternative compensation models can be explored, such as grant funding and research support. These models provide financial resources to support collaborative AI projects and facilitate the development of innovative solutions. Intellectual property rights and revenue-sharing arrangements can also be considered for successful AI tools, ensuring that benefits are distributed fairly among collaborators.

Seamless Integration of AI Tools into Existing Workflows

The challenges of integrating AI tools into existing workflows necessitate careful planning and effective change management strategies. Disruption of established routines and practices can hinder the adoption of AI solutions and breeds contempt — humans having to change their process for the sake of a tool meant to support them. Minimizing this disruption requires careful planning, change management strategies, and a deep understanding of current workflow. User experience and interface design are crucial considerations, as intuitive and user-friendly interfaces enhance the acceptance and usability of AI tools. Ensuring that AI tools can communicate and share data with existing healthcare systems is key to seamless integration.

AI tools must align with diverse workflows, necessitating a user-centered design approach, adaptability, and rigorous real-world testing. A user-centered design approach ensures that the needs and preferences of end-users, such as physicians and healthcare professionals, are prioritized. Customizability and adaptability are important features, allowing AI tools to align with diverse workflows and individual preferences. Rigorous testing and evaluation in real-world clinical settings assess the performance, reliability, and safety of integrated AI tools, ensuring AI solutions enhance the delivery of healthcare services.

Examples of Clinical Collaboration and Integration

AI in Alzheimer's Detection: In an early Alzheimer's detection model, AI development involves neurologists, data scientists, ethicists, and administrators, along with patients. Various incentives encourage active stakeholder participation.

AI in Emergency Triage: Suppose an AI tool is designed to streamline the patient triage process in emergency departments. Successful integration requires not just accuracy and reliability but also user-friendly design for healthcare professionals operating under stress. This involves integration with existing electronic health record systems, customization according to hospital protocols, and comprehensive training and support for medical professionals.

Conclusion

Despite slowing the process and being resource-intensive, collaborative approaches and workflow integration are vital for responsible AI development in healthcare. Involving diverse stakeholders, fostering partnerships, providing fair compensation, and ensuring seamless integration into existing workflows promotes trust and acceptance, paving the way for transformative advancements in healthcare. If these principles are ignored, we risk developing a final tool that does not actually serve the needs of patients or requires such a different and time-intensive new workflow that the tool will not be adopted by its users.

Rule 7: Regulatory Compliance and Safety

William Small, MD, MBA and Raouf Hajji, MD, PhD

AI has the potential to improve the accuracy of disease diagnosis and outcome predictions, enhance research equity, streamline the healthcare workflow, save cost and time, and personalize medical education. However, as we discuss in this charter, AI in healthcare comes with significant risks of errors and patient harm, risk of bias and increased health inequalities, and vulnerability to hacking and data privacy breaches. Given these concerns, it is crucial to implement robust safety measures to protect patients by adhering to regulatory guidelines without damaging their promising advances. Thankfully, efforts have been made in the United States and European Union to establish regulations that protect the rights of anyone impacted by AI systems. However, the unique, rapidly-evolving nature of LLMs pose the urgent need for new regulations because of their rapid integration into healthcare systems worldwide without a coherent global strategy.

The Current Regulatory Approach to AI in 2023

Regulatory guidelines for AI are currently developing, but they lag behind AI's rapid progress. However, both the European Commission and the U.S. White House have made significant efforts. The European Commission has been setting legal standards for AI safety, while the U.S. White House has been working on establishing an AI Bill of Rights.

The EU's AI Act of 2021 proposes a new regulatory framework. This framework targets the safe use of high-risk AI systems, including those aimed at improving health and safety. It also addresses systems whose failures could threaten fundamental human rights, such as personal data protection and freedom from discrimination. Developers of high-risk AI systems, particularly those related to healthcare, have certain responsibilities. They need to use high-quality data to train their models and maintain detailed records. They must be transparent about the AI's role and accuracy to both users and regulators. When necessary, they need to provide human oversight and maintain strong cybersecurity measures.

Failure to meet these requirements could have serious consequences. It can put individual providers and health systems at legal risk and threaten patients' fundamental rights. These rights are comprehensively laid out in the Blueprint for an AI Bill of Rights, distributed by the White House in 2023.

The Blueprint comprises five key principles, and by aligning with regulatory frameworks, developers of high-risk AI systems can mitigate most risks in implementing these systems into healthcare. The relevance of each of the 5 principles to healthcare is as follows:

Safe and Effective Systems have extensive (and recorded) pre-testing with high-quality data and independent evaluations, risk mitigation plans, and maintenance efforts. LLMs are currently trained with publicly available data; to be deployed in healthcare, they should be validated by healthcare professionals and/or trained specifically on healthcare data, such as Google's Med-PALM 2.

Algorithmic Discrimination Protections are essential to the long-term success of any developer of a high-risk AI system and health equity. Models should include regular equity assessments, which include using data representative of the people the systems are intended for. Regulators should enlist leaders of health equity and clinical informatics to ensure biases that exist in the data are not carried forward into future models.

Data Privacy's importance has been ingrained in US healthcare workers since 1996's HIPAA mandate; informed consent, ethical usage and distribution of data, and agency of the individuals whose data is being used are all principles that can be easily translate into this new era involving the development and deployment of AI systems in healthcare.

Notice and Explanation is an extension of informed consent that states AI systems should make all users aware that an automated system is being used to aid decision-making, using plain language in explanations. This is key in healthcare settings, where the tasks are often complex and require translation into language understandable to those with limited health literacy.

Human Alternatives, Consideration, and Fallback acknowledge the necessity for individuals to opt out from interacting with an AI system. End-users (whether doctors or patients) should have an accessible human alternative or fallback approach. Additionally, human oversight of AI systems is required, and they should be staffed by teams of individuals with the proper technical and healthcare expertise. There must be strict regulations on the types of outputs offered without oversight, such as explanations of their personal health data given by a chatbot..

The US Food and Drug Administration (FDA) has been a pioneer in the domain of AI regulation for healthcare settings. They achieved this by categorizing software as a medical device (SaMD) and expanded their existing regulatory framework to encompass devices incorporating machine learning (ML) and AI. Additionally, they introduced a Predetermined Change Control Plan. This mandates the creators of novel technologies to foresee potential risks and future modifications to their software, and detail their approach for output monitoring, risk mitigation, and technology adaptation.

Their 2021 SaMD Action Plan aims to address concerns that surfaced from their initial approach. The first step is to develop a **tailored regulatory framework**. This begins with a draft guidance on an updated Predetermined Change Control Plan and then collects feedback from the community to bolster the safety of novel AI/ML algorithms in healthcare. This should incorporate requirements for backup safety measures, including human oversight and fail-safe mechanisms to safeguard patients. Special attention is also warranted for algorithms that assist with diagnoses or medical procedures.

The next item of action is the establishment of **Good Machine Learning Practice (GMLP)**. This is targeted to be accomplished via participation in both local and global organizations. The advent of GMLP will necessitate robust cybersecurity, and collaboration with the FDA Medical Device Cybersecurity Program is essential for this. Additionally, we recommend developing training materials to educate clinicians who will be using these systems for patient care.

The third action from the FDA, a **patient-centered approach incorporating transparency to users**, describes the FDA plans to conduct a public workshop. This workshop will focus on developing trust in AI/ML-based devices among users, all the while acknowledging the complexity of these algorithms.

Their fourth action item is to establish **regulatory science methods related to algorithm bias and robustness**. The aim is to identify and eradicate discrimination by supporting and collaborating with researchers at their Centers for Excellence in Regulatory Science and Innovation (CERS

Lastly, to effectively evaluate **Real-World Performance (RWP)**, the FDA underscores t necessity of prospective data collection and monitoring. They also intend to solicit public feedback for this last action item. Through RWP monitoring, manufacturers of SaMDs incorporating AI/ML can promptly address safety concerns and gather feedback from end-users.

By collaborating with the public and cutting-edge researchers, the FDA aims to construct an effective regulatory framework for integrating AI/ML into SaMDs. This should provide guidance on how health systems can utilize and regulate these algorithms, whether they are incorporated into novel medical devices or used independently.

Large Language Models Pose Unique Regulatory and Safety Challenges

Current LLMs, trained on billions of parameters and serving as all-purpose AI models, have enormous potential to facilitate clinician workflows, enhance accurate diagnoses, and educate patients about their medical care. However, they pose unique regulatory and safety challenges because of the data that are used to train them, their ability to teach themselves based on data used to prompt text generation, and the potential for humans to rely on machine output without careful validation. Additionally, LLMs are known to "hallucinate," generating factually incorrect text to fulfill the goal of the prompt. While the current regulatory approaches to AI are a good start, regulators must take further steps to account for the variety of unique challenges posed by newer LLMs, especially as they advance to include images, video, and document inputs.

Healthcare institutions have been racing to use LLMs in their care delivery despite the safety and regulatory concerns. Mesko and Topol report a list of 10 unique regulatory challenges related to LLMs, many of which overlap with this document. Several that we do not specifically address are worth mentioning, including intellectual property and data ownership.

Addressing the unique challenges presented by LLMs is critical for ensuring their safety within the healthcare sector. The EU has already taken steps, for example proposing new copyright regulations for generative AI, necessitating developers disclose any copyrighted materials used in training their models. While some may suggest halting the development of advanced models until appropriate regulations are set, such a proposition is unlikely to succeed due to its anti-competitive nature, the evident economic benefits of the technology, and the inherent self-improvement design of the models.

Looking ahead, we foresee a wave of proposals and draft regulations from governments and research agencies worldwide, which seems to be the most effective route towards standardization and safety. We recommend that these new measures build upon existing frameworks that protect the rights of individuals interacting with AI models. Furthermore, these regulations should specifically address each of the unique challenges posed by LLMs, as described above. This approach will ensure a balanced, safe, and beneficial integration of LLMs into healthcare settings.

Conclusion

This is an extraordinary moment in the history of healthcare with a great opportunity to build the foundation for safe and ethical use of transformative AI, including generative LLMs. While the potential exists for great improvements in diagnostics, efficiency, health equity, cost savings, and personalized medical education for clinicians and patients, each example also poses risks to each of its users, including medical errors, discrimination, privacy, and transparency. A global effort is required among regulators, researchers, and end-users to identify and mitigate the idiosyncratic risks of these models and build robust regulatory guidelines that encapsulate each of them.

A massive effort is underway to address the unique challenges posed by generative AI. The lack of a unified global strategy highlights the nascency of research around the use of these models in the real world. Due to the rapid progress of these models and the potential for inclusion of even more personal data such as biometrics and videos, regulators, administrators, and end-users must be dynamic in how they develop, implement, monitor, and adapt these models. For responsible and safe implementation of AI in healthcare, the fundamental human rights outlined above must be considered with every regulation and alteration.

Rule 8: AI Education & Training

Sarah Gebauer, MD and Carly Eckert, MD, MPH

Many of the clinical use cases for AI intersect with patient care, and therefore will affect the workflows of physicians. As we write this in 2023, most physicians are unprepared to use and interact with AI in these settings. Specifically, they are underprepared to interrogate AI, appropriately critique or advocate for its use, and explain the technology to patients. Just like a new medication, procedure, or device — physicians must be trained and educated in how each AI tool can help — and what the potential side effects or complications may be.

As AI tools become more common in healthcare, physicians will be looked to for leadership in their governance. This includes establishing and overseeing frameworks related to acceptable use cases, data rights, and mitigating model biases, to name a few. Comprehensive education and training programs are required to support healthcare providers as their roles evolve.

These educational needs extend across a wide spectrum - from seasoned clinicians to students entering medical schools and even premedical courses, where initiatives around data knowledge and computational thinking are being taught.

In 2020, a survey found that 44% of practicing physicians felt that their medical education did not prepare them for new technologies, and only 7% felt prepared in 2020 to use artificial intelligence in clinical settings. These discrepancies point to the need for comprehensive education for physicians related to evaluation of clinical effectiveness, interpretation of underlying AI assumptions and processes, ethical considerations including bias mitigation, and evaluation of privacy and security. In this chapter, we explore the role of physicians in healthcare AI and the importance of comprehensive education and training programs.

Evaluation of Clinical Effectiveness

For physicians to effectively determine whether AI tools are appropriate for use in clinical settings, they should be able to evaluate their effectiveness. This involves understanding the mechanisms of how the tool works, the data used, the cohort upon which the tool was developed, and the logic behind how it arrived at its output.

Clinicians need to be aware of AI development to be able to answer critical questions, such as “Where was this built?”, “What data was it trained on?” and “What metrics are most meaningful for the use case?” (These same questions are asked when reviewing medical research.)

Just as a physician might question a lab value that seems inconsistent with a patient’s clinical state, they must also learn to question the AI tools at their disposal. Physicians must also learn to effectively interact with their institution’s data science, engineering, and IT teams, as these personnel are essential collaborators for AI tools implementation. Finally, physicians should understand how a tool arrived at its conclusions.

Consequences of AI Tools Implementation and Resource Allocation

Resource allocation is a key component of healthcare - some patients may receive an intervention while others do not. Additionally, physicians must learn to ask “who might this harm” when considering the use of an AI tool and have the agency to refuse to use an AI tool when the potential for harm is present.

Ethics and Bias Mitigation

While we cover Ethics and Bias Mitigation in Rule 2, we think it is critical that this is called out in our section on education and training as well. Physicians must be trained not only on AI tools — but they must be specifically educated on potential bias due to AI training sets. Additionally, physicians themselves will likely have ethical concerns about AI tools, and will also need help learning how to talk about AI tools with patients, just as they have had to learn about discussing new medications or procedures with patients as well.

Patient Concerns, Including Privacy and Security

Patients may have concerns about the privacy and security of their health data when used in conjunction with AI tools. Physicians should be aware of these concerns and address them.

Furthermore, patients may be uneasy about the potential of AI tools to replace human decision making. Physicians should be able to explain how AI tools are used and why they are important without causing undue concern. They may need to prepare for cases where patients ask that their data not be used for AI development or that their care is not guided by AI. Talking points related to data security, privacy, and applied technologies may become as necessary as the consent process that physicians are familiar with today.

Conclusion: A Framework for Educating Physicians

We propose a systematic and agile framework to educate physicians and the AI adjacent skills necessary. We can take some lessons from the implementation of electronic health records (EHRs), which is the most recent experience for most physicians learning a new technology. The general principles learned from the introduction of EHRs include the following:

Physician champions play an important role in the implementation of new technology in healthcare. However, they need support to be successful. This support could include decreased clinical responsibilities to devote appropriate time to the implementation process, immediate technical assistance when needed, and robust and tailored training modules.

Hands-on, immediate assistance with new technologies is essential for their successful implementation. This assistance can help physicians overcome any challenges that they may face during the implementation process and ensure that the technology is used effectively.

Developing a plan for the implementation, use, and feedback of technology that includes significant physician input is important to ensure that the technology is used effectively. Physicians have unique insights into how technology can be used in clinical settings, and their input can help ensure that the technology is tailored to their needs.

Just-in-time training: AI models and tools should provide education and background that can be rapidly reviewed at the time of use, like Duke's nutrition label concept for models or MDCalc's "Pearls and Pitfalls" supportive content for its digital tools.

Rule 9: Patient-Centered Outcomes and Value in Healthcare

Graham Walker, MD

The advent of AI in healthcare will foster a paradigm shift towards personalized (it will be possible), efficient (it will be faster), and data-driven (the data will exist) patient care. However, the crux of healthcare innovation lies not at all with technological advancement — but entirely with ensuring that these advances lead to tangible improvements in patient-centered outcomes and true healthcare value. AI deployment should be oriented towards enhancing clinical outcomes that patients care about, reducing over-diagnosis and over-treatment, and improving the value of healthcare.

Understanding Patient-Centered Outcomes and Value in Healthcare

Patient-centered outcomes refer to the outcomes that patients care about — that truly impact their lives: did they live or die? Did they get admitted to the hospital or experience a complication? Patient-centered outcomes are a primary focus of evidence-based medicine, and we believe this emphasis must continue with AI. Patient-centered outcomes should be the **primary** metrics by which the success of AI implementations is measured.

Healthcare value, on the other hand, encapsulates the quality of health outcomes achieved per dollar spent. AI tools should either improve outcomes at the same cost or maintain outcomes at lower cost. This must include the reduction of over-diagnosis/treatment, which leads to unnecessary costs and potential harm to patients. We already see these challenges today with more sensitive testing, diagnosing clinically insignificant (and patient-unimportant) disease.

But I lowered the blood pressure! Or: Why This Matters

While many patients (and doctors) may think "lowering the blood pressure by 30 points" is a great achievement, this is only because we know that long-term control of hypertension correlates with patient outcomes of lowered risks of heart attack, stroke, or death. AI tools cannot focus on fascination or novelty or simple numeric or surrogate measures of health.

By focusing on patient-centered outcomes, AI applications can help the fundamental objective of medicine: to enhance the health and well-being of human beings. These tools must additionally consider all outcomes of interest — positive and negative — and weigh these against each other: Yes, perhaps we lower the patient's risk of stroke by anticoagulating them, but how much did we increase their risk of life-threatening hemorrhage?

Examples and Recommendations in Clinical Practice

Patient-Important Outcomes: Consider an AI tool that can evaluate patients with osteomyelitis, and researchers target antibiotic duration. The AI model is able to reduce oral antibiotic duration by 3 hours. This model should not be implemented into clinical practice as a 3-hour reduction in antibiotic necessity is probably not of importance to most patients, but a shorter hospital stay (or prevention of hospital admission altogether) would be.

Healthcare Value: An AI system designed to detect pulmonary emboli considers the clinical significance of the detected emboli (clot burden, hemodynamic stability, vital signs, cardiac biomarkers), including considering that some emboli may be clinically insignificant or even false positive studies. By incorporating the risk of anticoagulation and PE treatment into its model, the system can help ensure that its decisions prioritize patient-important outcomes rather than mere detection. (The PERC Criteria follow this risk:benefit analysis today.)

Conclusion

AI design and implementation must stay true to the ethos of medicine, prioritizing patient-centered outcomes and healthcare value, just as evidence-based medicine (EBM) has done as well. While the potentials of AI are vast, these principles provide a critical foundation, ensuring that AI applications in healthcare effectively enhance patient care, reduce healthcare inefficiencies, and contribute to overall healthcare improvement. By remaining anchored in these objectives, AI could truly revolutionize the landscape of healthcare, moving us toward a future of more personalized, efficient, and patient-oriented medical care, but we must make sure its target is outcomes that matter to humans.

Rule 10: Understanding the Limits of AI

S. Morgan Jeffries, MD, and Graham Walker, MD

Many physicians are extremely excited about what is possible as AI is integrated into healthcare delivery — including the authors of this very charter. We imagine the promise of unparalleled advances in diagnostics, therapeutics, patient care, and clinical decision making. AI may bring us back — or at least closer — to the bedside, when the EHR has pulled us so far away. But make no mistake, we are cautiously — not blindly — optimistic, and it is critical that we end this roadmap by acknowledging AI's limitations and where over-reliance on AI tools could harm patients. (It's also a nice reminder of all the incredible abilities of the human mind.)

Strengths and Weaknesses of AI

AI is undoubtedly a powerful tool. It can process immense amounts of data at unmatched speeds, find patterns that are invisible to the human eye, and perform multiple tasks in parallel while humans tend to need to work in serial fashion. These abilities could improve diagnostic accuracy, prevent medical error, facilitate predictive modeling, and personalize patient care. However, these capabilities are not without boundaries, and it is essential to be aware of these limitations — and why AI always needs to be the assistant, the co-pilot, or the consultant to the physician. Bear in mind that these are moving targets; as capabilities expand, some of these limitations may change.

1. The Limits of AI Autonomy

AI cannot function autonomously (nor do we believe it should in medical care). As automated tools improve, though, learned carelessness leads to increased automation bias, whereby users come to trust automated tools uncritically. This wouldn't be a problem if these tools were always right, but automated systems (including AI) occasionally fail; depending on the stakes, this can be disastrous. Humans — yes, including physicians — are more prone to these biases when multitasking or otherwise under increased cognitive load. Clinicians must stay vigilant, must be educated about automation bias, and must influence the development of AI to minimize risks going forward.

2. Machines Learn What the Data Teaches Them

AI, specifically machine learning, relies on very large quantities of data for training. It can only detect patterns present in its training data and none outside of that. This means that it will tend to pick up biases present in the training data and that its performance will suffer if real world patients behave differently than training data. (We should note that humans may be susceptible to these pitfalls as well.) A key difference is that humans engage in continuous learning, whereas models must be retrained explicitly. For more information, see Rule 3's Types and Instances of Bias and Rule 5's Shifts and Drifts and What To Do When There's a Problem.

3. AI Does Not Learn from Its Mistakes

Another inherent limitation is that AI does not learn from its mistakes in the same way humans do. As humans, we instinctively learn from errors and strive to avoid making the same mistakes again. Conversely, ML models operate in two modes: training, where their settings are repeatedly improved and optimized, and inference, where they do actual predictive work. The models a practicing physician will encounter will be in inference mode, and models in inference mode do not learn from mistakes. Given the same data, an AI tool in inference mode would produce the same error on its thousandth run as on its first. This can only be addressed by retraining the model, highlighting the need for regular AI model evaluation and updates.

This might not square with one's experience with AI chatbots, seemingly capable of learning mid-conversation. The secret is that chatbots are built using language models, and the models can "remember" past responses, allowing them to provide updated responses over time. However this is not infinite; for example if you start a new conversation, the model will behave as though the original conversation never happened.

4. AI Does Not Ruminate

Current AI systems lack the ability to ponder over problems, an inherent characteristic in the human thought process. A patient presenting with a perplexing or unclear combination of signs and symptoms often causes physicians to think deeply and mull over the case, even after the interaction is over. AI does not exhibit this behavior. AI, by default, operates in what Daniel Kahneman termed as 'fast' thinking mode. While it is possible to induce Large Language Models (LLMs) to 'think' more slowly through careful prompting, inconsistencies need to be addressed by the user explicitly.

5. All AI Is Narrow

You may have heard the term artificial general intelligence (AGI), which describes an AI that's capable of solving a broad range of problems. This contrasts with artificial narrow intelligence (ANI), also referred to as narrow AI, which is strictly limited in scope. Currently, all AI is quite narrow. A sepsis model can only predict sepsis. While generative language models have a much broader scope, they're still limited to producing text from text inputs. This could change over time, but there are no commercially available AI systems today that can process a broad range of sensory inputs or carry out the broad range of tasks that a human can. They cannot comprehend or interpret non-verbal cues, body language, or socio-economic contexts that are pivotal in healthcare scenarios. This limitation underscores the importance of integrating AI tools with human care, which can account for these non-verbal and socio-cultural factors.

6. AI Rarely Outperforms Experts Overall

While AI has made significant strides in many areas, its performance rarely exceeds human expertise in practice, for many reasons. AI, compared to human thought, is comparatively rigid. Humans can to adapt to changes in data distribution and modify their thought process on-demand.

But one can still make use of AI's more limited expertise by using it to complete tasks a human cannot easily do. AI models are tools that can enhance the decision-making process by providing data-driven insights, but they are not infallible. In the complex, multifaceted realm of healthcare, where every decision can significantly impact a patient's life, the expertise, judgment, and experience of human practitioners continues to be indispensable.

7. AI Can't Form Stable Relationships

AI today can produce one-off responses that are perceived as empathetic, but much more goes into forging and maintaining human relationships – even relatively one-sided relationships like the one between a patient and a physician. Current AI systems cannot process facial expressions, body language, or tone of voice. They also cannot mirror these expressions. Forming a relationship with another human entails recalling facts about each other, and developing an intuitive sense of values, and interests. A language model could approximate this by tracking the relevant information — but doing this well is nontrivial.

Examples in Clinical Practice

Radiology and AI Misinterpretation: An AI algorithm is designed to identify brain tumors from MRI scans. While it may be effective in diagnosing clear-cut cases, it could struggle with complex or atypical presentations — since extremely rare diseases would not be widely available in its training data (since they are by definition rare). A radiologist on the other hand can reference textbooks and educational materials and brings years of experience and nuanced understanding to the table, allowing them to detect unusual presentations or rule out false positives.

AI and Mental Health: Consider the use of AI in mental health. AI algorithms might be able to identify patterns or key words indicative of conditions like depression, based on text analysis from patient interactions. However, these tools might miss crucial elements of the patient's mental state that are communicated nonverbally or via subtle cues in spoken language, a limitation that highlights the irreplaceable value of human providers.

AI in Patient Triage: An AI system in an emergency department may be trained to triage patients based on data from initial assessments. However, a patient's condition can quickly change, which may not be immediately reflected in the AI's data. A human clinician's ability to assess and reassess in real-time is crucial, and this nuanced, dynamic decision-making capability is something AI systems do not possess.

Conclusion

While AI can augment human capabilities in the medical field, human expertise and judgment are irreplaceable. In complex decision-making scenarios, human clinicians provide nuanced understanding and flexible decision-making ability that AI systems lack. AI in healthcare, therefore is a tool to assist — not replace — healthcare providers.

The capabilities of AI complement the human touch in healthcare. Clinicians don't just diagnose and treat—they empathize, reassure, and build trust. AI can support these processes by providing efficient and reliable data analysis, but it cannot replace the human connection and trust that is integral to the healing process.

In conclusion, the future of healthcare AI comes with our responsibility to understand its limitations. Recognizing these will guide us in implementing AI responsibly, as a co-pilot or assistant to human expertise, mitigating potential risks, and maximizing benefits. By striking a balance between AI assistance and human intervention, we can ensure the delivery of optimal care and improved patient outcomes.

05. The Physicians' Charter Conclusion

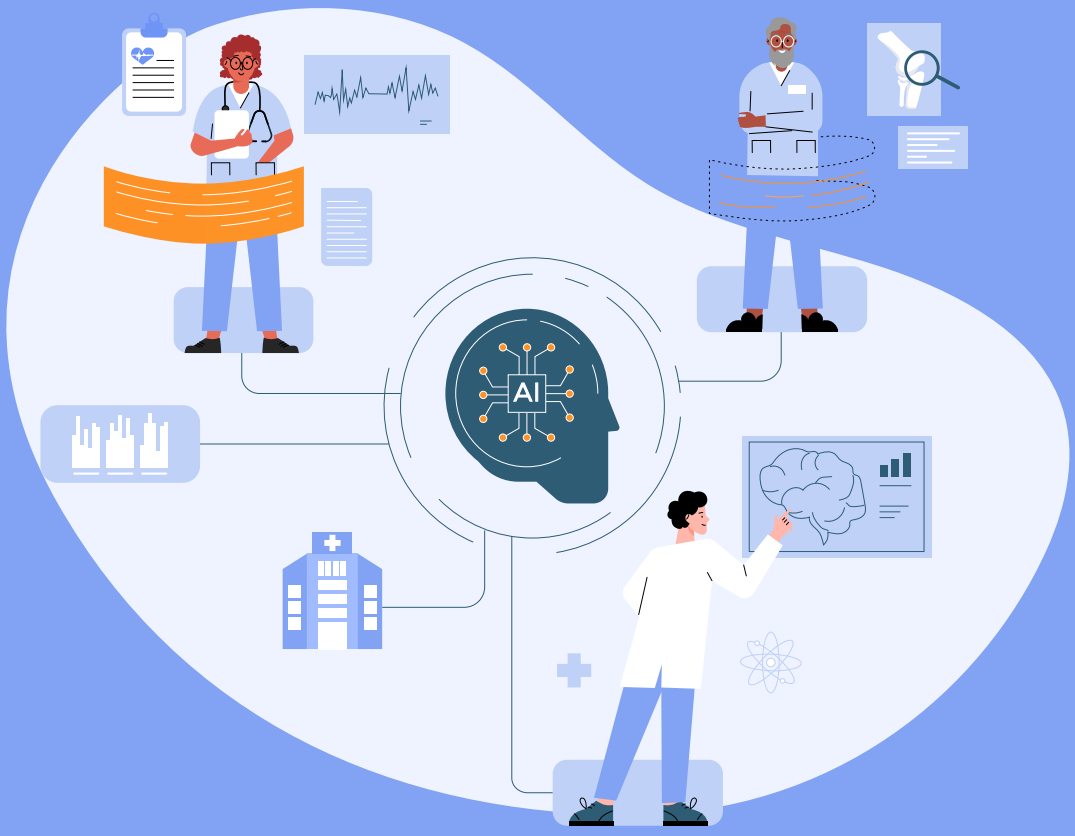
As we conclude our guide, a new era dawns — one brimming with potential. Just as antibiotics and X-rays have improved the practice of medicine, so too will artificial intelligence. AI's potential to redefine the practice of medicine for patients, physicians, healthcare workers, policymakers, and ethicists is enormous. There are countless ways that it could help improve diagnostics, deliver better education, offer new treatment recommendations, and even allow physicians more time with their patients. Yet it brings with it a critical obligation to ensure ethical, safe, and respectful deployment — and physicians are essential experts who can help navigate this space.

And as we delve deeper into this brave new world, it has become clearer that introducing AI into healthcare is not a mere plug and play scenario. Rather, it requires continuous, meticulous effort in calibrating, validating, and updating the AI systems, which in turn requires time and resources for ensuring the reliability, safety, accuracy, and trustworthiness of these tools.

Throughout this document, we've discussed our 10 Rules of the Road for AI Implementation, each one serving as a beacon to guide us through the complexities of incorporating AI into healthcare. But as we write these rules in the summer of 2023, we acknowledge the need for adaptability. Rather than rigid edicts, they are dynamic guideposts, capable of evolving in step with the expanding and transformative world of AI. Our framework, therefore, must be as fluid and adaptable as the technology it seeks to shepherd. And when in doubt, our values must serve as our compass, steering our decisions and molding the development and implementation of AI.

The trust given to us by patients is the sacred bedrock upon which all healthcare is built. It is equally susceptible to erosion from indifference or malpractice as it is from unethical or misguided AI deployments. Missteps — no matter the intent — have the potential to echo far beyond individual patients, beyond the confines of one doctor or one institution, and beyond the boundaries of medical specialties. These reverberations could indeed stall or even halt further advancement of AI in healthcare.

AI offers immense promise, from improved patient outcomes to enhanced provider experiences. Let us embrace AI as a valuable ally and co-pilot — supporting the patient-physician relationship — and enabling us to help, counsel, and guide our patients through their lives. Our journey towards responsible AI usage is a path to better healthcare for all. We invite you to join us on this journey — guided by firm principles yet adaptable in our approach, and always anchored by our dedication to our patients.



06. About the Authors

Anthony Cardillo, MD is a pathologist and Clinical Informatics fellow at NYU Langone. His primary interests are in medical cybersecurity and the digital transition of pathology. He presently serves on two national committees involving artificial intelligence and ethics in the College of American Pathologists and the American Medical Informatics Association. More recently, Dr. Cardillo was recognized in The Pathologist's Power List in 2021 and 2022, and in 2023 placed in the US and UK Summit for Democracy competition to develop secure AI models.

William "Will" Collins, MD is a hospital medicine physician and Clinical Assistant Professor at the Stanford School of Medicine. He is also the current president of the Society of Hospital Medicine San Francisco Bay Area Chapter. He has been captivated by both the potential and the risk of AI applications in medicine. From his experience in clinical research, he is interested in designing rigorous trials to assess AI interventions to show meaningful outcomes for patients and medical providers.

Dustin Cotliar, MD MPH, brings significant care delivery expertise that comes from over eight years of clinical practice and studying healthcare policy and management at Columbia University. He has served as a clinical consultant with the Kaiser Family Foundation where he published health system research that has been cited in articles by the NY Times, VOX, politico, and others. A recent first-place winner at MIT's Hacking Medicine, one of the largest clinical hackathons in the country, Dr. Cotliar is passionate about building innovative clinical products, especially those rooted in artificial intelligence and machine learning.

Carly Eckert, MD, MPH is a physician technologist located in Chapel Hill, NC. She is double-boarded in preventive medicine and clinical informatics. Carly has led clinical and data teams within healthcare startups for nearly a decade. Her areas of focus include AI governance, ethics, and bias. She also enjoys teaching physicians and other healthcare providers on the topics of practical and applied AI solutions and how to communicate with technical teams.

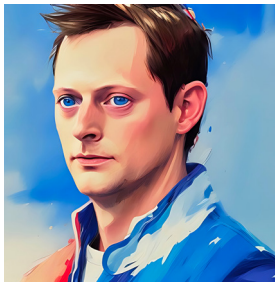
Sarah Gebauer, MD is an experienced hospital leader and healthcare technology consultant with a background in clinical informatics. She's passionate about physician engagement with artificial intelligence and founded Machine Learning for MDs, a free online community providing education, training, and networking for physicians in the AI space.

Raouf Hajji, MD, PhD is an Assistant Professor of Internal Medicine, Medicine Faculty of Sousse, Tunisia. With his expertise in clinical practice, biomedical research, and academia, he is the author, reviewer, and editor of many peer-reviewed medical journals and book chapters. He is Co-founder and Medical Lead of International Medical Community (IMC), an international initiative working as an Innovation Health technologies Hub with the main scope of advancing international cooperation and creating a link to cutting-edge technologies for the healthcare sector worldwide. You can join him on LinkedIn where he publishes weekly medical newsletter: Healthcare Present & Future with Updates on Biomedical Research, Academia, Clinical Practice and Emerging Technologies in Healthcare.

S. Morgan Jeffries, MD is a neurohospitalist and physician informatician at Geisinger, where his work focuses on quality measures, workflow improvements, and AI strategy. He's also an assistant professor at the Geisinger Commonwealth School of Medicine and a member of Epic's Adult Neurology Specialty Steering board. He's interested in the similarities and differences between human and AI minds, AI safety and alignment, and AI evaluation. He occasionally writes on LinkedIn and less frequently on X (née Twitter).

Matt Sakumoto, MD is a virtualist primary care physician in San Francisco, and Adjunct Clinical Professor at UCSF focusing on virtual care and clinician efficiency tools for the EHR. With prior industry experience at multiple telehealth startups and as a clinician-advisor to many early-stage companies, he is passionate about exploring and expanding the digital health landscape.

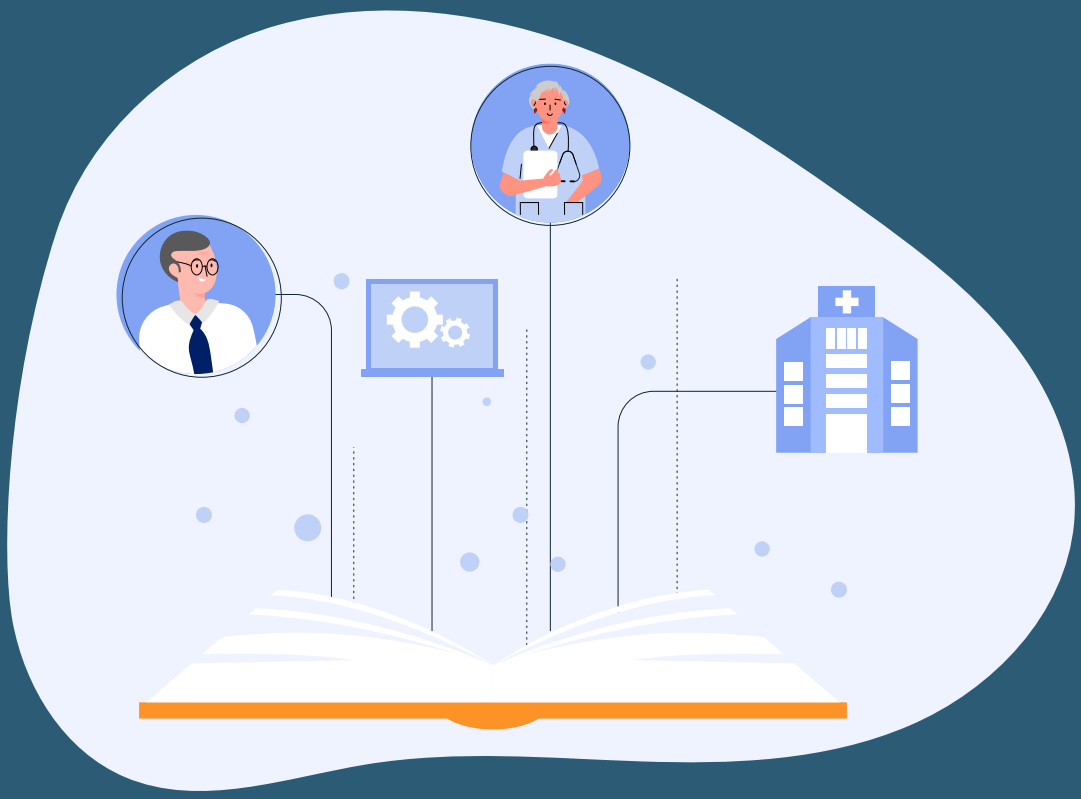
William Small, MD, MBA is a hospital medicine physician and clinical informatics fellow at NYU Langone Health who is focused on the impact of communication technologies on the clinician experience with the EHR and patient outcomes. He is dedicated to understanding how best to evaluate outputs of generative AI and is a key member of the team evaluating the effects of generative AI chatbot integration into patient-provider Inbasket communications on provider efficiency and satisfaction.



Graham Walker, MD is an emergency physician and clinical informaticist in San Francisco, California with The Permanente Medical Group (TPMG) and enjoys working at the intersection of technology and medicine and created and organized the Physicians' Charter for Responsible AI. He also built MDCalc and theNNT, two free online resources that have allowed millions of clinicians from around the world to incorporate evidence-based decision-making into their medical practice. You can find him at LinkedIn writing about medicine and technology and referring to himself in the third person.

Disclaimer

The views and opinions expressed by the authors do not necessarily reflect the views of our employers or other organizations with which we're affiliated. While every effort has been made to ensure the accuracy and currency of the information, the rapidly evolving nature of AI in healthcare means that some details may change over time; we hope to review this document annually. This document was created by individual physicians in our free time. © August 17, 2023.



07. Acknowledgements

- Many thanks first to Dr. Scott Campbell, an emergency physician, founding member of the American Board of AI in Medicine, and incredibly gifted AI expert who helped develop the concept for this charter in the first place.

- Thank you to Dr. Sarah Gebauer for creating the Machine Learning for MDs online community, without which this paper would have struggled to find diverse authorship, as well as Dr. Shoreh Irani for her Physician-Led AI in Medicine Facebook group.

- Appreciation to Drs. Basil Kahwash, Jane Wang, Oluseyi Fayanju for their edits and suggestions during the writing process, along with Dr. Sarah Gebauer and Dr. Raouf Hajji for their review and editing as well. Graham Walker's parents (Susan and Don Walker) also provided editing advice, as they have been doing ever since Graham was old enough to write.

- Thank you to Drs. Edward Yap and Charulata Ramaprasad for providing their opinions as the project was just starting out.

- Gratitude to Dr. Justin Norden for his leadership in the intersection of AI and medicine, and for providing feedback about the charter as it developed.

- We look forward to other organizations and partnerships building on and collaborating with this charter — and recommend that we continue to work with expediency in this area, as the world of AI is moving extremely fast, and at a much faster pace than medicine would normally progress. We were inspired by the Coalition for Health AI (CHAI) at https://www.coalitionforhealthai.org/ and the Health AI Partnership at https://healthaipartnership.org/. We also look forward to the work of the National Academy of Medicine's Health Care Artificial Intelligence Code of Conduct, and hope it will start to provide expedited recommendations as well.

Finally, this document was born out of the discussions the physician creators of MDCalc, both hopeful-yet-concerned after seeing the power of generative AI in late 2022. MDCalc supported and this project's work and allowed it to go from vague concept to illustrated, edited, completed execution.

08. Further Reading and Supporting Research

Introduction, Values, Mission, and Vision

a. This is a wonderful piece on Dr. Rene Theophile Hyacinthe Laënnec, a French physician and the inventor of the stethoscope in the early 1800s.

b. In a similar vein, we have supporting histories of the PET scan and medical ultrasound.

c. Stanford's Human-Centered Artificial Intelligence has a great brief overview of different AI terms and definitions.

1. Human-Centered Design and Engagement

a. The New England Journal of Medicine provides a thorough summary of Artificial Intelligence and Machine Learning in Clinical Medicine, including history, ML, and chatbots, with proposals for research standards as well.

b. Stanford's HAI (Institute for Human-Centered Artificial Intelligence) is an outstanding resource for all-things AI, and its Values section aligns well with this Charter's vision, and its Humanity section aligns perfectly with our first chapter.

c. This paper from San Diego comparing ChatGPT responses to Reddit physician responses made headlines and sparked controversy when it suggested that OpenAI's tool provided more empathetic responses than the physician users on Reddit.

d. Want to learn more about Human-Computer Interaction? Ben Shneiderman's book entitled Human-Centered AI is a great place to get started.

e. And for more HCI information, look to HCI International's book series and conference.

2. Data Quality and Privacy

a. This 2023 article from Computers in Biology and Medicine delves into the barriers of AI adoption in healthcare, focusing on various privacy and data concerns, and presents an overview of advanced privacy-preserving techniques like Federated Learning and Hybrid Techniques.

b. A team in China provides an excellent review on federated learning and privacy-preserving algorithms as solutions to data fragmentation and privacy challenges in healthcare AI.

c. This is a summary of a roundtable discussion by the US Department of Health and Human Services (HHS) on the opportunities, challenges, and strategies for using data to train AI models in healthcare, offering recommendations for HHS and stakeholders to further AI advancements.

3. **Ethics and Bias Mitigation**
    a. This NEJM paper discusses the use of race in predictive algorithms, the problems that arise when using race, and highlights the importance of knowing what goes into algorithms.
4. **Trust: Transparency, Explainability, and Accountability**
    a. Carnegie Mellon's Violet Turri has an outstanding piece on "What is Explainable AI?"
    b. This famous paper from Microsoft on explainable models revealed an issue with a neural network that was predicting that patients with asthma had a lower likelihood of mortality from pneumonia (when in actuality they have a higher mortality); the model was making technically accurate conclusions, but made these conclusions because asthmatic patients were more often managed in the ICU, lowering their mortality due to more aggressive, intensive care.
    c. This paper from PLOS is an outstanding review of the ethical, theoretical, and practical concerns around AI models and tools — specifically focusing on how emergency dispatch operators did not adopt a tool that predicted which emergency calls were for a cardiac arrest case because they did not trust or understand it.
    d. Epic's Sepsis model is discussed in this paper and is unfortunately a good example of a model failing "in the wild."
5. **Continuous Validation, Monitoring, and Improvement**
    a. This Lancet paper suggests concerns around generalizability of models in healthcare and explains the reasons that models may not be as generalizable as we would like to think.
    b. This NEJM correspondence (in particular, its Table 1) provides an overview of approaches to recognizing and addressing dataset shift.
6. **Collaborative Approach and Workflow Integration**
    a. Authors from Ohio State provide a roadmap for the integration of AI into Radiology workflows specifically, from the Journal of Medical Imaging.
    b. European Radiology reviews the challenges — and offers solutions to them — in this piece, again focusing around AI in radiology.
    c. This article from the UK discusses advances in Human-Computer Interaction, breaking the paper up into 6 categories: Interfaces, Visualization, Electronic Health Records, Devices, Usability, and Clinical Decision Support Systems.

7. **Regulatory Compliance and Safety**
    a. The FDA provides guidance for AI and ML in Software as a Medical Device applications.
    b. The FDA also has a helpful navigator to help developers determine if their software is a medical device.

8. **Education and Support**
    a. This paper from Health Education UK argues that the healthcare workforce in the UK will need education and training — and the creation of an educational framework — to use AI successfully.
    b. This article interviews 45 physician champions and discusses what they felt were critical to the adoption of a new EHR and what challenges they faced.
    c. Health Affairs discusses 7 lessons from EHR implementation — including hands-on training.
    d. Here are 10 more lessons learned from an academic medical center that adopted an EHR for its 6 hospitals, 2 campuses, and 46 outpatient sites.

9. **Patient-Important Outcomes and Value in Healthcare**
    a. This paper reviews the very concept of patient-important outcomes, and acknowledges that medicine doesn't often ask patients what's important to them as an outcome.
    b. Even in research today, we don't focus nearly enough on patient-important outcomes — in diabetes and critical care as just two examples.

10. **Understanding the Limits of AI**
    a. Thinking, Fast and Slow is a book by psychologist Daniel Kahneman, who describes two systems that humans use when thinking; a fast, instinctive system, and a slow, deliberative thought process.
    b. The complementarity-driven deferral to clinicians (CoDoC) system proposes a model that could even help clinicians decide when to rely on AI tools and when to defer to clinician judgment.
    c. This paper demonstrates how AI can be helpful to humans — by re-ordering CT scan reading queues — without replacing physician interpretation.
    d. The tragic crash of flight AF447 is an example of the devastating consequences of automation bias.
    e. Automation bias is hard to overcome, even when humans are educated and warned that it exists.