Health Care AI: Law, Regulation, and Policy.

W. Nicholson Price II

University of Michigan Law School, wnp@umich.edu

Available at: https://repository.law.umich.edu/book_chapters/357

Follow this and additional works at: https://repository.law.umich.edu/book_chapters

Part of the Health Law and Policy Commons, and the Science and Technology Law Commons

Publication Information & Recommended Citation

Artificial Intelligence in Health Care

The Hope, the Hype, the Promise, the Peril

Michael Matheny,
Sonoo Thadaney Israni, Mahnoor Ahmed,
and Danielle Whicher, Editors
This publication has undergone peer review according to procedures established by the National Academy of Medicine (NAM). Publication by the NAM signifies that it is the product of a carefully considered process and is a contribution worthy of public attention, but does not constitute endorsement of conclusions and recommendations by the NAM. The views presented in this publication are those of individual contributors and do not represent formal consensus positions of the authors’ organizations; the NAM; or the National Academies of Sciences, Engineering, and Medicine.

Library of Congress Control Number: 2020938860

Copyright 2022 by the National Academy of Sciences. All rights reserved.

Printed in the United States of America

INTRODUCTION

As discussed in previous chapters, artificial intelligence (AI) has the potential to be involved in almost all aspects of the health care industry. The legal landscape for health care AI is complex; AI systems with different intended uses, audiences, and use environments face different requirements at state, federal, and international levels. A full accounting of these legal requirements, or of the policy questions involved, is far beyond the scope of this chapter. Additionally, the legal and regulatory framework for AI in health care continues to evolve, given the nascent stage of the industry.

In this chapter, we offer an overview of the landscape through early April 2019 and undertake three tasks. First, we lay out a broad overview of laws applicable to different forms of health care AI, including federal statutes, federal regulations, and state tort law liability. Second, we address in considerable depth the regulatory requirements imposed on AI systems that help inform or make decisions about individual patients, such as diagnosis or treatment recommendations; these systems are referred to in this report as clinical AI. Clinical AI faces the closest scrutiny, especially by the U.S. Food and Drug Administration (FDA) and by other regulatory agencies internationally. These systems must demonstrate safety and efficacy. They may also generate liability under state tort law, which performs its own regulatory role and is intimately tied to the way FDA oversees clinical AI systems. Third, we note the legal and policy issues around privacy and patient data that affect clinical AI as well as other health care AI systems. Throughout the chapter, we highlight key challenges, opportunities, and gaps in the current framework. The chapter concludes with key considerations for addressing some of these issues.
OVERVIEW OF HEALTH CARE AI LAWS AND REGULATIONS IN THE UNITED STATES

Developers and users of health care AI systems may encounter many different legal regimes, including federal statutes, federal regulations, and state tort law. Below are a few of the most significant among these laws and regulations:

- **Federal Food, Drug, and Cosmetic Act (FDCA):** FDA enforces the FDCA, which regulates the safety and effectiveness of drugs and medical devices, including certain forms of medical software (21 U.S.C. §§ 301 ff.). The bulk of this chapter describes the application of the FDCA to health care clinical AI systems.

- **Health Insurance Portability and Accountability Act (HIPAA):** In addition to the Privacy Rule (described in more detail below), HIPAA authorizes the U.S. Department of Health and Human Services to enforce the Security Rule (45 C.F.R. Parts 160 and 164). These rules create privacy and security requirements for certain health information. The HIPAA Breach Notification Rule also requires certain entities to provide notifications of health information breaches (45 C.F.R. §§ 164.400–164.414). To the extent that the development or use of health care AI systems involves health information covered by HIPAA, those requirements may apply to developers or users of such systems.

- **Common Rule:** The Common Rule sets requirements for research on human subjects that either is federally funded or, in many instances, takes place at institutions that receive any federal research funding (45 C.F.R. Part 46). Among other things, most human subjects research must be reviewed by an institutional review board (45 C.F.R. § 46.109). These requirements can apply to AI used for research or the research used to create health care AI. The Common Rule is enforced by the Office for Human Research Protections.

- **Federal Trade Commission Act (FTCA):** The FTCA prohibits deceptive and unfair trade practices affecting interstate commerce (15 U.S.C. §§ 41–58). These could include acts relating to false and misleading health claims, representations regarding a piece of software’s performance, or claims affecting consumer privacy and data security. Health care AI products may raise any of these types of claims. The Federal Trade Commission (FTC) enforces the requirements of the FTCA.

- **FTC Health Breach Notification Rule:** This FTC rule, separate from HIPAA’s Breach Notification Rule, requires certain businesses to provide
notifications to consumers after a breach of personal health record information, including information that may be collected to train, validate, or use health care AI systems (16 C.F.R. Part 318). The FTC enforces this rule.

- **State tort law:** When one individual or entity injures another, tort law may allow the injured individual to recover damages. Injury could result from the use of health care AI systems, including when the behavior of developers, providers, hospitals, or other health care actors falls below the standard of care. State law determines the applicable standard of care and when tort liability will exist.

We summarize each of these categories of regulatory and legal oversight by application in Table 7-1, referencing the applicable laws and regulations for different types of AI systems. Liability refers to the legal imposition of responsibility for injury through the state tort law system.

### TABLE 7-1 | Typical Applicability of Various Laws and Regulations to U.S. Health Care Artificial Intelligence Systems

<table>
<thead>
<tr>
<th>Clinical Urgency</th>
<th>Type of health care AI system</th>
<th>Description</th>
<th>HIPAA/Office of Civil Rights</th>
<th>FTC</th>
<th>FDCA</th>
<th>Liability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Research</td>
<td>AI intended to assist human subjects research</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Operations</td>
<td>AI that is used to enhance clinical operations, such as patient management, scheduling, and physician documentation</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>General health and wellness</td>
<td>AI that is used by consumers for entertainment</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Clinical: mobile engagement; health and wellness; medical device data systems</td>
<td>AI that is used by consumers for entertainment AI in certain categories for which FDA has announced that it does not intend to enforce FDCA requirements</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Direct-to-consumer</td>
<td>AI that is marketed directly to consumers and constitutes a medical device</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Clinical: informing clinical management</td>
<td>AI that assists physicians by informing, enhancing, aggregating, and verifying information</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Clinical: driving clinical management</td>
<td>AI that assists physicians by giving treatment, diagnosis, or screening advice, while relying on physician interpretation of said advice to direct patient care</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Clinical: treating or diagnosing</td>
<td>Autonomous AI that provides treatment or diagnoses, screens, disease without physician interpretation</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>
SAFETY AND EFFICACY OF CLINICAL SYSTEMS

A key set of laws work to ensure the safety and efficacy of medical technology, including clinical AI systems. The principal requirements are determined by the FDCA and enforced by FDA. State tort law also plays a role in ensuring quality by managing liability for injuries, including those that may arise from insufficient care in developing or using clinical AI.

The raison d’être of clinical AI systems is to be coupled with and to inform human decision making that bears upon the content and conduct of clinical care, including preventive care, to promote favorable, equitable, and inclusive clinical outcomes and/or mitigate risks or interdict adverse events or nonoptimal outcomes. Regulatory authorities in various countries, including FDA, expect the pharmaceutical, medical device, and biotechnology industries to conduct their development of all diagnostics and therapeutics (including companion and complementary diagnostics and therapeutics) toward the goal of safer, more efficacious, and personalized medicine. This development should result in care that is, at a minimum, not inferior to conventional (non-AI-based) standard-of-care outcomes and safety endpoints. Health services are expected to fund such AI-coupled diagnostics and therapeutics, and prescribers and patients are, over time, likely to adopt and accept them. Increased development of “coupled” products (including clinical AI systems) could result in “safer and improved clinical and cost-effective use of medicines, more efficient patient selection for clinical trials, more cost-effective treatment pathways for health services,” and a less risky, more profitable development process for therapeutics and diagnostics developers (Singer and Watkins, 2012).

The right level of regulation requires striking a delicate balance. While the over-regulation or over-legislation of AI-based personalized medical apps may delay the translation of machine learning findings to meaningful, widespread deployment, appropriate regulatory oversight is necessary to ensure adoption, trust, quality, safety, equitable inclusivity, and effectiveness. Regulatory oversight is also needed to minimize false-negative and false-positive errors and misinterpretation of clinical AI algorithms’ outputs, actions, and recommendations to clinicians. Recent examination of the ethics of genome-wide association studies for multifactorial diseases found three criteria necessary for identification of genes to be useful: (1) the data in the studies and work products derived from them must be reproducible and applicable to the target population; (2) the data and the derived work products should have significant usefulness and potential beneficial impact to the patients to whom they are applied; and (3) the resulting knowledge should lead to measurable utility for the patient and outweigh associated risks or potential harms (Jordan and Tsai, 2010). Thus, regulatory standards for clinical
AI tools should at least extend to accuracy and relevancy of data inputs and model outputs, marketing of AI systems for specific clinical indications, and transparency or auditability of clinical AI performance.

Medical Device Regulation

Some AI systems, particularly those algorithms that will perform or assist with clinical tasks related to diagnosis, interpretation, or treatment, may be classified as medical devices and fall under applicable FDA regulations. Other AI systems may instead be classified as “services” or as “products,” but not medical devices (see Box 7-1). FDA’s traditional regulatory processes for medical devices include establishment registration and listing plus premarket submissions for review and approval or clearance by FDA’s Center for Devices and Radiological Health Office of Device Evaluation or Office of In Vitro Diagnostics and Radiological Health. In the United States, the Medical Device Amendments of 1976 (P.L. 94–295) to the FDCA (21 U.S.C. § 360c) established a risk-based framework for the regulation of medical devices. The law established a three-tiered risk classification system based on the risk posed to patients should the device fail to perform as intended. The FDCA (21 U.S.C. § 360j) definition of a medical device is summarized in Box 7-1.

**BOX 7-1**


*Medical Device Definition*

[A]n instrument, apparatus, implement, machine, contrivance, implant, in vitro reagent, or other similar or related article, including a component part, or accessory which is:

[One from the following]

Recognized in the official National Formulary, or the United States Pharmacopoeia, or any supplement to them OR intended for use in the diagnosis of disease or other conditions, or in the cure, mitigation, treatment, or prevention of disease, in man or other animals OR intended to affect the structure or any function of the body of man or other animals, and which does not achieve its primary intended purposes through chemical action within or on the body of man or other animals AND which does not achieve its primary intended purposes through chemical action within or on the body of man or other animals and which is not dependent upon being metabolized for the achievement of its primary intended purposes.
The 21st Century Cures Act (Cures Act, P.L. 114-255) was signed into law on December 13, 2016. The significant portion with regard to clinical AI systems is Section 3060 (“Regulation of Medical and Certain Decisions Support Software”), which amends Section 520 of the FDCA so as to provide five important exclusions from the definition of a regulatable medical device. Under Section 3060 of the Act, clinical decision support (CDS) software is nominally exempted from regulation by FDA—that is, it is defined as not a medical device—if it is intended for the purpose of:

(i) displaying, analyzing, or printing medical information about a patient or other medical information (such as peer-reviewed clinical studies and clinical practice guidelines);
(ii) supporting or providing recommendations to a health care professional about prevention, diagnosis, or treatment of a disease or condition; and
(iii) enabling such health care professional to independently review the basis for such recommendations that such software presents so that it is not the intent that such health care professional rely primarily on any of such recommendations to make a clinical diagnosis or treatment decision regarding an individual patient.

This exemption does not apply to software that is “intended to acquire, process, or analyze a medical image or a signal from an in vitro diagnostic device or a pattern or signal from a signal acquisition system” (21st Century Cures Act § 3060). FDA has stated that it would use enforcement discretion to not enforce compliance with medical device regulatory controls for medical device data systems, medical image storage devices, and medical image communications devices (FDA, 2017a). The 21st Century Cures Act codifies some of FDA’s prior posture of restraint from enforcement.

Under this system, devices that pose greater risks to patients are subject to more regulatory controls and requirements. Specifically, general controls are sufficient to provide reasonable assurance of a Class I device’s safety and effectiveness, while special controls are utilized for Class II devices for which general controls alone are insufficient to provide reasonable assurance of device safety and effectiveness (21 C.F.R. § 860.3). FDA classifies Class III devices as ones intended to be used in supporting or sustaining human life or for a use that is of substantial importance in preventing the impairment of human health, or that may present a potential unreasonable risk of illness or injury, and for which insufficient information exists to determine whether general controls or special controls are sufficient to provide reasonable assurance of the safety and effectiveness of a device (21 C.F.R. § 860.3). This highest risk class of devices is subject to premarket approval to demonstrate a
reasonable assurance of safety and effectiveness. Even for this highest risk class of devices, the evidence FDA requires for premarket approval has long been flexible, varying according to the characteristics of the device, its conditions of use, the existence and adequacy of warnings and other restrictions, and other factors. There is generally more flexibility in the amount of clinical evidence needed for medical devices than for drugs and biological products, because they are subject to different statutory criteria and the mechanism of action and modes of failure are generally more predictable and better characterized for devices than for drugs and biological products.

Additionally, the design process for a medical device is more often an iterative process based largely on rational design and non-clinical testing rather than clinical studies. However, this last aspect is not, in general, true for clinical AI systems. The machine learning process is itself a kind of observational research study. In some cases—particularly for medium- and high-risk clinical AIs—the design process may depend on lessons learned as such tools are deployed or on intermediate results that inform ways to improve efficacy (FDA, 2019b). The Clinical Decision Support Coalition and other organizations have recently opined that many types of clinical AI tools should not be regulated or that the industry should instead self-regulate in all application areas that FDA chooses not to enforce on the basis of their review of risks to the public health. Notably, the principles and risk-based classification processes have recently been updated to address requirements for software as a medical device (SaMD) products (see FDA, 2017c § 6.0, p. 11; IMDRF, N12 § 5.1).

It is worth noting the distinction between CDS software tools, including clinical AIs, that replace the health professional's role in making a determination for the patient (i.e., automation) and those that simply provide information to the professional, who can then take it into account and independently evaluate it (i.e., assistance). The former may be deemed by FDA to be a medical device and subject to medical device regulations. Under the 21st Century Cures Act, if a CDS product has multiple functions, where one is excluded from the definition of a medical device and another is not, FDA can assess the safety and effectiveness to determine whether the product should be considered a medical device (21st Century Cures Act § 3060). Also, FDA can still regulate the product as a medical device if it finds that the software “would be reasonably likely to have serious adverse health consequences” or meets the criteria for a Class III medical device. Clinical AI systems that are deemed to be medical devices will generally require either De Novo or premarket approval submissions (FDA, 2018a). In some instances, where a valid pre-1976 predicate exists, a traditional 510(k) submission may be appropriate.

Note, too, that the 21st Century Cures Act’s statutory language, while already in force, is subject to implementing regulations to be developed by FDA over
time and leaves considerable ambiguity that subjects developers of clinical AI systems to FDA enforcement discretion. For example, uncertainty remains when software is being used in “supporting or providing recommendations,” or when it “enables a health care professional to independently review the basis for [its] recommendations.” FDA has issued some draft guidance (FDA, 2017b), and more guidance will undoubtedly be forthcoming. But ambiguity will likely be present nonetheless, as will the possibility of enforcement discretion.

Oversight of safety and effectiveness does not just come from regulators, whether domestic or international. In particular, diagnostic testing that is provided by laboratories and other enterprises as services is subject to oversight provided by the Clinical Laboratory Improvements Act of 1988 (CLIA, P.L. 100-578) and the Patient Safety and Quality Improvement Act of 2005 (P.L. 109-41). Certain clinical AI tools that are services rather than products may be appropriate to regulate under CLIA. It is possible that some clinical AIs—especially ones that have aspects similar to diagnostics classified as laboratory-developed tests (LDTs), developed and performed in university-based health facilities or other provider organizations—may be deployed strictly as services for patients in the care of those institutions and not marketed commercially.

**FDA’s Digital Health Initiative**

FDA has expressed interest in actively promoting innovation in the digital health space. FDA’s proposed Digital Health Software Precertification (Pre-Cert) Program aims to (1) substantially reduce regulatory burdens for most suppliers and operators of clinical AI systems and (2) improve the health system’s rates of responsiveness to emerging unmet health needs, including personalized medicine (FDA, 2018c).

The 21st Century Cures Act and FDA documents reflect an increasing realization that data from real-world operations are necessary for oversight. Health care information technology (IT) systems are so complex and the conditions under which clinical AI systems will operate so diverse that development, validation, and postmarket surveillance must depend on utilizing real-world data and not just clinical trials data or static, curated repositories of historical data. “Real-world data (RWD) are data relating to patient health status and/or the delivery of health care routinely collected from a variety of sources,” including electronic health record (EHR) systems (FDA, 2019c). “Real-world evidence (RWE) is the clinical evidence regarding the usage, and potential benefits or risks, of a medical product derived from analysis of RWD” (FDA, 2019c). All of these are subject to applicable HIPAA and other privacy protections such that RWD and RWE must be rigorously de-identified (El Emam, 2013) prior to use for the secondary
purposes of clinical AI development and productization. RWE and RWD are discussed in greater detail below.

The goal of FDA’s Pre-Cert Program (FDA, 2019a) is to establish voluntary, tailored, pragmatic, and least-burdensome regulatory oversight to assess software developer organizations of all sizes. The Pre-Cert Program simultaneously aims to establish trust that developers have adequate quality management system (QMS) processes in place and a culture of quality and organizational excellence such that those developers can develop and maintain safe, effective, high-quality SaMD products. The Pre-Cert Program leverages the transparency of organizational QMS compliance and product safety as well as quality metrics across the entire life cycle of SaMD. It uses a streamlined premarket review process and leverages postmarket monitoring to verify the continued safety, effectiveness, and quality performance of SaMD in the real world. The premarket review for a precertified organization’s SaMD product is informed by the organization’s precertification status, precertification level, and the SaMD’s risk category. With this program, FDA envisions leveraging the risk-category framework for SaMD developed by the International Medical Device Regulators Forum (IMDRF) to inform the risk category (FDA, 2017c, 2019b). The IMDRF framework describes the spectrum of software functions, some of which may not meet the definition of a device in Section 201(h) of the FDCA and others that may meet the definition of a device, but for which FDA has expressed that it does not intend to enforce compliance. For the purposes of the Pre-Cert Program, the application of FDA’s long-established risk category framework would remain consistent with the current definition of device under Section 201(h) of the FDCA and FDA’s current enforcement policies. The IMDRF framework establishes types and subtypes of SaMD products based on the state of the health care condition and the significance of the information provided by the products (IMDRF, 2014).

Most clinical AI systems are multielement “ensembles” of a plurality of predictive models with an evidence-combining “supervisor” module that establishes a collective answer or output from the ensemble-member models’ execution. Clinical AI involves prediction, classification, or other intelligence-related outputs. These are generated from data supplied as inputs to the model, from fewer than 10 to many hundreds of phenotypic input variables or—in the case of time-series or spectrum-analytic AI systems, image-processing AI systems, or AI systems that include genomics biomarkers—a large number of engineered features that are derived from very high-dimensional raw data inputs. Likewise, present-day genomics-based diagnostics typically involve dozens of input variables, for which there are regulatory policies and procedures that have been established for more than 10 years. These govern existing regulated diagnostics, such as in vitro diagnostic multivariate index assays (IVDMIAs) (FDA, 2007a).
Not all clinical AI systems will manifest hazards or have risk levels comparable to those associated with existing IVDMA products. However, the methodology, review, and clearance criteria that have been found effective for the regulation of IVDMAAs may form a useful point of reference for the regulatory practice for clinical AI systems.

Clinical AI Systems That May Merit Different Regulatory Approaches

FDA has indicated that it will apply a risk-based assessment framework, where the risk level of different clinical AI systems will be influenced by the different types of on-label clinical indications and contexts in which they are intended to be used, plus the different situations in which their off-label usage might plausibly be anticipated, adopting the IMDRF framework (FDA, 2019b; IMDRF, 2014).

For example, a clinical AI system's intended use might be as a screening test to determine the person’s susceptibility to, or propensity in the future to, develop a clinical condition or disease that has not yet materialized; this affords time for longitudinal observation, repeat testing, and vigilance to monitor signs and symptoms of the emergence of the disease and is accordingly lower risk. Similarly, an AI system designed to classify a condition’s stage or current severity, or to establish the prognosis or probable clinical course and rate of progression of a condition, functions essentially like a biomarker that characterizes risk and does so in a manner that is amenable to multiple repeat tests and observations over a period of time.

Such situations have low time sensitivity and a plurality of opportunities for the experienced clinicians to review, second-guess, and corroborate the recommendations of the screening clinical AI system. In IMDRF parlance, these are clinical AI systems that “inform” clinical management but do not “drive” clinical management. Indeed, the “informing care” function of some present-day clinical AI tools of this type is to automatically/reflexively order the appropriate standard-of-care confirmatory diagnostic testing and monitoring. These clinical AI systems provide additional evidence or advice (e.g., regarding the likelihood of the condition screened for and/or the cost-effectiveness of pursuing a diagnostic workup for the condition) and promote consistency, relevancy, and quality in diagnostic workups. In general, such screening or informational clinical AI systems will be classified as having low risk. As such, many clinical AI systems are outside the formal scope of medical device regulation and do not require establishment registration and listing or other regulatory filings (21st Century Cures Act § 3060).

By contrast, some classification, forecasting, and prognostic biomarker clinical AI algorithms that instead drive clinical management and/or involve clinical indications may be associated with a medium or high risk; the AI systems could
contain faults that cause harm via commissive or omissive errors, either directly or through clinicians’ actions or inaction. Perioperative, anesthesiology, critical care, obstetrics, neonatology, and oncology use-cases are examples of medium- or high-risk settings (Therapeutic Monitoring Systems, Inc., 2013). In such situations, there is great time sensitivity and there may be little or no time or opportunity to seek additional testing or perform more observations to assess the accuracy of the AI’s recommendation or action. In some instances, such as oncology and surgery, the decision making informed by the AI tool may lead to therapeutic actions that are not reversible and either close other therapeutic avenues or alter the clinical course of the illness and perhaps its responsiveness to subsequent therapy. Such AI tools would, by Section 3060 of the 21st Century Cures Act, be formally within the scope of medical device regulation and would require establishment registration, listing, and other regulatory filings—De Novo, 510(k), premarket approval, or precertification—and associated postmarket surveillance, reporting, and compliance procedures.

**Explainability and Transparency from a Regulatory Perspective**

AI systems are often criticized for being black boxes (Pasquale, 2016) that are very complex and difficult to explain (Burrell, 2016). Nevertheless, such systems can fundamentally be validated and understood in terms of development and performance (Kroll, 2018; Therapeutic Monitoring Systems, Inc., 2013), even if not in terms of mechanism—and even if they do not conform to preexisting clinician intuitions or conventional wisdom (Selbst and Barocas, 2018). Notably, the degree of “black box” lack of explainability that may be acceptable to regulators validating performance might differ from the amount of explainability clinicians demand, although the latter is an open empirical question. This chapter addresses explainability to clinicians and other nonregulators only to the extent that it interacts with regulatory requirements. Instead, the focus is largely on validation by regulators, which may be satisfied by some current development processes.

While the rest of this section focuses on how explainability and transparency may or may not be required for regulators to oversee safety and efficacy, regulators may also require explainability for independent reasons. For instance, regulators may require clinical AI tools to be explainable to clinicians to whose decision making they are coupled; to quality assurance officers and IT staff in a health provider organization who acquire the clinical AI and have risk-management/legal responsibility for their operation; to developers; to regulators; or to other humans. The European Union’s General Data Protection Regulation (GDPR) right to explanation rules, for instance, enacted in 2016 and effective
May 2018, applies to AI systems as well as humans and web services (Kaminski, 2019) and governs European Union citizens worldwide. Similar standards may be implemented in the United States and other jurisdictions. Such standards and regulations are important for public safety and for the benefits of clinical AI systems to be realized through appropriate acceptance and widespread use. However, the notion of explainability is not well defined. There is a lack of agreement about both what constitutes an adequate explanation of clinical AI tools, and to whom the explanation must be provided to conform to applicable right to explanation rules and thus be suitable for regulatory approval.

Current right to explanation regulations and standards fail to acknowledge that human data scientists, clinicians, regulators, courts, and the broader public have limitations in recognizing and interpreting subtle patterns in high-dimensional data. Certain types of AI systems are capable of learning—and certain AI models are capable of intelligently and reliably acting upon—patterns that humans are entirely and forever incapable of noticing or correctly interpreting (Seblst and Barocas, 2019). Correspondingly, humans, unable to grasp the patterns that AI recognizes, may be in a poor position to comprehend the explanations of AI recommendations or actions. As noted, the term “black box” is sometimes pejorative toward AI, especially neural networks, deep learning, and other fundamentally opaque models. They are contrasted to logistic regression; decision-tree; and other older-technology, static, deterministic models—all with low dimensionality but are able to show the inputs that led to the recommendation or action, with variables that are generally well known to the clinician and causally related.

If society, lawmakers, and regulatory agencies were to expect every clinical AI system to provide an explanation of its actions, it could greatly limit the capacity of clinical AI developers’ use of the best contemporary AI technologies, which markedly outperform older AI technology but are not able to provide explanations understandable to humans. Regulators do not currently require human-comprehensible explanations for AI in other industries that have potential risks of serious injury or death. For example, autonomous vehicles are not required to provide a running explanation or commentary on their roadway actions.

While requiring explainability may not always be compatible with maximizing capacity and performance, different forms of transparency are available that might enable oversight (see Figure 7–1). For instance, transparency of the initial dataset—including provenance and data-processing procedures—helps to demonstrate replicability. Transparency of algorithm or system architecture is similarly important for regulatory oversight. When AI systems are transparent not just to the regulator but more broadly, such transparency can enable independent validation and oversight by third parties and build trust with users.
AI Logging and Auditing

Today, developers creating clinical AI systems with their enterprises’ risk-management and product liability exposures in mind are engineering and testing their clinical AI deliverables with Agile (Jurney, 2017) or other controlled software development life cycle (SDLC) methods. Defined, well-managed, and controlled SDLC processes produce identifiable and auditable systems and maintain controlled documents of the systems’ development processes under the developers’ written, reviewed, and approved standard operating procedures. They conform to QMS principles (see ISO–9001, ISO–13485, and 21 C.F.R. Part 820), FDA device master record type, Current Good Manufacturing Practices (CGMPs), and applicable laws and regulations. These include design assurance, design control, hazard analysis, and postmarket surveillance (21 C.F.R. Part 822) provisions. Such industrial-strength developers of clinical AI systems also engineer their systems such that the systems’ operation creates (1) a persistent, archived log of each transaction or
advisory output that each clinical AI system performs; (2) the versioning of the AI system’s elements that performed the transaction, traceable to the data sources; and (3) the validation and software quality-assurance testing that led to the AI systems being authorized for production and subsequent use. These logs enable the examination of the inputs, outputs, and other details in case of anomalies or harms. The logs are open to the clinician-users, employers, organizations who acquire/authorize the AI system’s deployment (e.g., health provider organizations, health plans, or public health agencies), regulators, developers, and the courts. The individual release-engineered and version-controlled instances of present-day industrial-strength clinical AI systems are identifiable and rigorously auditable, based on these SDLC controlled-document artifacts, which are maintained by the developer organization that owns the intellectual property.

For this type of clinical AI system, the regulatory agencies’ traditional submissions and compliance processes for SaMD are feasible and may not need substantial alteration (e.g., FDA, 2018b). The types of evidence required by plaintiffs, defendants, counsel, and the courts may not need substantial alteration, although the manner of distributed storage, retrieval, and other aspects of provisioning such evidence will change. Moreover, the availability of such evidence will not be significantly altered by the nature of clinical AI systems, provided that developers follow QMS and CGMPs and maintain conformity, including controlled-document artifacts retention.

Some clinical AI tools will be developed using RWD. Because RWD are messy in ways that affect the quality and accuracy of the resulting inferences, as described in Chapter 6, more rigorous requirements for auditing clinical AI systems developed and validated using RWD will need to be established.

**AI Performance Surveillance and Maintenance**

Several architectural and procedural aspects of machine learning–based clinical AI systems will require significant changes in regulatory compliance and submissions, in terms of scale and scope. In modern AI applications using dynamic data sources—such as clinical data streams stored in EHRs, data collected via sensor-enabled wearable devices and combined with other forms of data, and other RWD—the AI models and algorithms are likely to experience drift over time or as the algorithms are deployed across institutions whose catchment areas and epidemiology differ (dataset shift, see Quiñonero-Candela et al., 2009; Subbaswamy et al., 2019). These longitudinal drifts and shifts entail expanded design control, design assurance, and evidentiary requirements, as discussed in Chapter 6 and below. Therefore, the traditional approach of assessing performance using a static, limited dataset to make assessments about the ongoing safety of a system is inadequate with regard to
clinical AI. Continuous machine learning offers one solution for dataset shift and drift by updating with new population-specific data (FDA, 2019b). Not all clinical AI systems aim to do this, and very few clinical AI systems implement continuous learning today. The goals of learning health systems and personalized medicine do create an impetus for more continuous machine learning–based AI systems, as discussed in Chapters 3 and 6 in more detail. However, current regulatory and jurisprudential methods and infrastructure are not prepared for this.

**Natural Language Processing and Text Mining AI**

Unstructured notes constitute another important source of RWD, and appropriate standards for the extraction, parsing, and curation of unstructured information for clinical AI systems is therefore another open area requiring regulatory oversight. Natural language processing (NLP) algorithms and text mining are important for certain kinds of clinical AI that use unstructured data such as clinical impressions and other remarks, as discussed in Chapter 5.

There will be a need for retention and curation of the unstructured source-text documents as well as the discrete labels or concept codes and values derived by NLP from those documents. Retention of all of these is necessary because NLP algorithms may change over time. The underlying lexical reference data and parameters that govern the parser’s operation may likewise change from one release to the next. Thus, release engineering regression testing and validation of successive releases of a clinical AI model that depends on unstructured text must be able to demonstrate that the NLP subsystem continues to meet its specifications and delivers to the clinical AI model inputs that are substantially equivalent to the results it delivered for the same test cases and document content in previous releases. Furthermore, there is natural variability in how different individuals speak. Unlike physiology, factors such as culture and training affect how individuals describe a phenomenon. Clinical AI systems must be robust to these variations.

**Clinical Decision Support Systems**

Another architectural factor to consider when regulating clinical AI systems is that traditional CDS systems have tended to be embedded in tangible medical devices or in single-site on–premises EHR systems (Evans and Whicher, 2018). The system configurations and dated–signed records of changes in such architectures are readily auditable by users, regulators, and courts. By contrast, many contemporary AI systems are deployed on cloud-based, geographically distributed, nondeterministically parallelized, spatially arbitrary computing architectures that, at any moment, are physically unidentifiable. To create and maintain a full log of each processor that
Artificial Intelligence in Health Care

contributed in some part to the execution of a multielement ensemble model AI is possible in principle but would likely be cost-prohibitive and too cumbersome to be practical. Therefore, the limited traceability and fundamental non-recreatability and non-retestability of a patient’s or clinician’s specific execution of an AI system that may have contained a fault or that produced errors or failures—untoward, unexpected deviations from its specifications, validation testing, and hazard analysis—may pose particular problems for regulators, courts, developers, and the public. These nondeterministic, noncollocation aspects of contemporary cloud-based AI implementations contrast with traditional criteria for tracking product changes (e.g., 510(k) supplements or adverse event reporting systems).

Hazard Identification, Risk Analysis, and Reporting Recommendations for Safe Clinical AI Systems

Identifying hazards is a necessary step to support safe system design and operations. Identifying hazardous situations requires experts to carefully and thoroughly evaluate the system via one of several methods. Successful assurance of public safety rests on (1) identifying and analyzing all significant possible scenarios that could result in accidents of differing severity, and (2) devising and documenting effective means of mitigating the scenarios’ likelihood, frequency, and severity. Although hazard identification and quantitative risk assessment are important, risk management also depends on qualitative or subjective judgments (e.g., human observation, intuition, insight regarding processes and mechanisms of causation, creativity in anticipating human actions and psychology, and domain expertise). Each of these judgments introduces biases and chances of omissions. Thus, hazard identification should be a structured process.

Traditional modes of risk assessment and hazard analysis (e.g., hazard and operability study [HAZOP] or process hazard analysis [PHA]) that have been used in regulation of medical devices for decades can also be used for clinical AI systems (ISO, 2009, 2016). However, new hazard types related to geographic dispersity and the dynamic, nondeterministic execution of cloud-based clinical AI systems and machine learning mean that new risks must be evaluated and new mitigations must be devised, tested, and documented. Clinical AI systems may exhibit emergent properties that depend on the whole system as it evolves in time and are not specified or statically defined by the system’s parts or subsystems (Johnson, 2002; Louis and Nardi, 2018). This means that the safety of a clinical AI system, like an autonomous vehicle, cannot “solely be analyzed or verified by looking at, for example, a hardware architecture” or only one physical system instantiation (Bagschik et al., 2018). As a result, requirements “must be derived in a top-down development process which incorporates different views on a system at all levels” (Bagschik et al., 2018).
Regulatory agencies should require clinical AI developers to conduct iterative system testing on multiple physical instances of the system and with enough iterations to provide reasonable assurance of detecting faults and hazards from many sources. These could include (1) logic races; (2) nonbinding and nonexecution of worker agents on cloud-based servers; (3) variable or prolonged latencies of data ingestion of accruing clinical information on which an AI depends into noSQL repositories; (4) nonexistent or erroneous mappings of input and output variables utilized by the AI algorithms to do their work; (5) nonreceipt or system-mediated rejection or nonstorage or nondisplay of the AI’s output to the relevant user(s); and even (6) potentially automatic software updates (i.e., unsupervised updates of clinical AI systems into “live” production environments, where they immediately begin to affect decisions and might not undergo local review and approval first by the user–clinicians’ IT or quality assurance staff). Such an iterative testing requirement is consistent with FDA’s recently issued guidance on addressing uncertainty in premarket approval decision making (FDA, 2018a).

For continuous learning and other dynamic, adaptive, and nondeterministic aspects of clinical AI systems and the computing architectures on which they are implemented, developers and regulators could usefully look to risk-assessment and -management methods that have been successfully used for two decades in the chemical process industry and other continuous-process operations such as public utilities (Alley et al., 1998; Allocco, 2010; Baybutt, 2003; Bragatto et al., 2007; Chung and Edwards, 1999; Frank and Whittle, 2001; Hyatt, 2003; Nolan, 2011; Palmer, 2004; Paltrinieri and Khan, 2016; Reniers and Cozzani, 2013; Venkatasubramanian et al., 2000; Villa et al., 2016). Chemical plants, for instance, depend on the availability of public utilities such as water, and chemical plant failure analyses note that dependence. A clinical AI system’s developer could similarly list the complete set of utilities (e.g., ongoing access to users’ de-identified datasets, on which the AI system’s development and validation are based, plus the user’s production system and its data on which the AI’s runtime operation depends) that might affect a specific node’s operation, and assess and manage each of them.

Clinical AI Systems’ Prediction/Classification Effectiveness- and Utility-Related Performance

The accuracy, sensitivity, and specificity of clinicians’ judgment and of traditional diagnostics are measures against which clinical AI systems’ statistical performance must compare favorably. FDA has, for many years, set forth guidance on procedures for assessing noninferiority and superiority of new medical products (FDA, 2016a; Newcombe, 1998a,b). For many applications, the so-called
Number Needed to Treat (NNT) and Number Needed to Harm (NNH) are useful measures of population-level clinical utility of a therapeutic or a diagnostic (Cook and Sackett, 1995; Laupacis et al., 1988). A product (i.e., medication or medical device) or a health service (i.e., clinical intervention, procedure, or care process) that has a very high NNT value (>100) or that has a very low NNH value (<10) is unlikely to meet clinicians’ or consumers’ expectations of probable clinical benefit and improbable clinical harm.

The international CONsolidated Standards of Reporting Trials (CONSORT, 2010) and STARD (Standards for Reporting of Diagnostic Accuracy Studies; Bossuyt et al., 2003) initiatives pertain to the verification of diagnostic accuracy conforming to existing good clinical practice rules and guidelines (Steyerberg, 2010). While these initiatives are not focused on studies that aim to demonstrate diagnostic device equivalence, many of the reporting concepts involved are nonetheless relevant and applicable to clinical AI. The CONSORT guidelines aim to improve the reporting of randomized controlled trials, enabling reviewers to understand their design, conduct, analysis, and interpretation, and to assess the validity of their results (CONSORT, 2010). However, CONSORT is also applicable to observational, nonrandomized studies and AI derived from machine learning.

According to a 2007 FDA guidance document, 

FDA recognizes two major categories of benchmarks for assessing diagnostic performance of new qualitative [classificatory or binomial/multinomial predictive] diagnostic tests. These categories are (1) comparison to a reference standard (defined below), or (2) comparison to a method or predicate other than a reference standard (non-reference standard).

... The diagnostic accuracy of a new test refers to the extent of agreement between the outcome of the new test and the reference standard. We use the term reference standard as defined in STARD. That is, a reference standard is “considered to be the best available method for establishing the presence or absence of the target condition.” It divides the intended use population into only two groups (condition present or absent) and does not consider the outcome of the new test under evaluation.

The reference standard can be a single test or method, or a combination of methods and techniques, including clinical follow-up [by appropriately credentialed clinician experts]. If a reference standard is a combination of methods, the algorithm specifying how the different results are combined to make a final positive/negative classification (which may include the choice and ordering of these methods) is part of the standard. (FDA, 2007b)
In addition to the area under the receiver operating characteristic (AUROC) curve, it is also important to evaluate the sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) as part of regulatory assessment of clinical machine learning predictive models and AI. These additional statistical performance metrics take the prevalence of the predicted outcome into account (unlike the AUROC curve, which is independent of prevalence [Cook, 2008]), and therefore have greater clinical relevance. Past studies have shown that PPV and the AUROC curve have minimal correlation for risk prediction models (Goldstein et al., 2017). Conventional statistical measures of accuracy, sensitivity, specificity, PPV, NPV, AUROC and partial AUROC, and so forth are, and will remain, the principal guides for regulatory clearance and enforcement.

Analytical validation involves “establishing that the performance characteristics of a test, tool, or instrument are acceptable” (Scheerens et al., 2017); the relevant performance characteristics are described in Chapter 5 and are important for regulatory oversight as well as internal analytical validation. These characteristics validate the AI’s technical performance, but not its usefulness or clinical value. Beyond conventional statistical metrics for diagnostic medical devices and regulatory agencies’ de facto norms for these, the objectives of clinical validation testing of an AI tool are to quantitatively evaluate a variety of practical questions:

- How did the AI algorithm outputs inform or obfuscate clinical decision support recommendations?
- How often were AI system recommendations reasonable compared to local licensed peer clinicians addressing similar situations, according to expert clinicians?
- How often did attending clinicians or other staff accept the AI tool’s recommendations, and how often did they override or interdict the action or recommendation of the AI tool?
- How often were the AI tool’s recommendations or actions unsafe or inefficacious, how often did they lead to errors or harm, and are the AI-associated rates of harm or nonbenefit unacceptably worse (i.e., statistically and clinically inferior) to what competent humans’ results are?

Furthermore, clinical utility is an inherent consideration for clinical AIs, as described in Chapter 6. “The conclusion [is] that a given use of a medical product will lead to a net improvement in health outcome or provide useful information about diagnosis, treatment, management, or prevention of a disease. Clinical utility includes the range of possible benefits or risks to individuals and populations” (FDA–NIH Biomarker Working Group, 2016).
Chapter 5 describes bias in sensitivity and specificity estimates in some detail. According to FDA,

sensitivity and specificity estimates (and other estimates of diagnostic performance) can be subject to bias. Biased estimates are systematically too high or too low. Biased sensitivity and specificity estimates will not equal the true sensitivity and specificity, on average. Often the existence, size (magnitude), and direction of the bias cannot be determined. Bias creates inaccurate estimates.

[Regulatory agencies hold that] it is important to understand the potential sources of bias to avoid or minimize them [Pepe, 2003]. Simply increasing the overall number of subjects in the study will do nothing to reduce bias. Alternatively, selecting the “right” subjects, changing study conduct, or data analysis procedures may remove or reduce bias. (FDA, 2007b)

These steps are essential to eliminate validation leakage and help to estimate the stability of the model over time.

Two main biases are important to consider: representational bias and information bias (Althubaiti, 2016). Representational bias refers to which individuals or data sources are represented in the data and which are not. Information bias is meant to represent collectively “all the human biases that distort the data on which a decision maker [relies] and that account for the validity of data [that is, the extent these represent what they are supposed to represent accurately]” (Cabitza et al., 2018). These two biases and the related phenomenon of information variability together can degrade the accuracy of the data and, consequently, the accuracy of the clinical AI model derived from them.

Real-World Evidence, Postmarket Surveillance, and Measurement of Clinical AI Systems’ Functional Performance

As regulators consider how to set the appropriate balance between regulatory oversight and access to new AI technology, some options include shifting the level of premarket review versus postmarket surveillance for safety and efficacy. Enhanced postmarket surveillance presents an attractive possibility to allow more streamlined premarket review process for AI technology and reflects the likelihood of more frequent product changes over time (FDA, 2019b). Given suitable streams of RWE, clinical AI systems are likely to learn on the fly from an ongoing data stream because population characteristics and underlying models can change. This requires the availability of high-quality labeled RWE as well as continuous oversight via postmarket surveillance, because unlike for a more
traditional FDA-approved medical device, many clinical AI systems will change over time with the addition of new data (FDA, 2019b), though for some models and developers the cost of adaptation may exceed the benefits. Especially for lower risk software, allowing market access on the basis of less substantial data on safety and efficacy and then monitoring carefully as the software is deployed in clinical practice may lead to smoother oversight that is still robust. However, prior efforts to rely on increased postmarket surveillance have encountered difficulty in developer compliance and agency enforcement (Woloshin et al., 2017), although the extent of this difficulty is contested (Kashoki et al., 2017).

As part of its Digital Health Innovation Action Plan, FDA is developing the Pre-Cert Program, in which certain developers can apply to be precertified based on a “robust culture of quality and organizational excellence” and commitment to monitoring real-world performance (FDA, 2019c). In the program as envisioned, precertified companies will be able to market lower risk SaMD without premarket review and will receive a streamlined premarket review for higher risk SaMD. FDA will work with developers to collect and interpret real-world information to ensure that the software remains safe and effective in the course of real-world use (FDA, 2019c), including the potential for updates and changes without further review (FDA, 2019b).

**Companion Diagnostic Versus Complementary Diagnostic Clinical AIs**

Diagnostics that inform the use of drugs, biologics, or therapeutic devices come in several regulatory forms. A companion diagnostic is sometimes required for drug/biologic/device approval (FDA, 2016b); the in vitro diagnostic device (IVD) and the associated therapeutic (i.e., drug, biologic, or other intervention) must be cross-labeled, and the IVD is thereafter used as a “gating” criterion for prescribing the therapeutic product. A complementary diagnostic, in contrast, merely provides additional information relevant to, or supplementary to and corroborative of, decisions guiding care of the patient in regard to the associated therapeutic product. Complementary diagnostics are not required for FDA approval of the associated therapeutic product, and need not be cross labeled. Finally, a combination product is a product composed of two or more regulated components produced and marketed as a single entity (21 C.F.R. § 3.2(e)). It is likely that many clinical AI systems whose hazard analyses indicate that they have medium or high risk could be successfully regulated as complementary diagnostic medical devices.

Two additional types of diagnostics are not regulated as commercially marketed products. An LDT is a type of IVD that is designed, manufactured, and used...
within a single health services facility for the care of patients for whom named clinicians in that facility have responsibility. Diagnostic tests that are not marketed commercially beyond the therapeutics development process, clinical trials, and regulatory marketing approval are generally referred to by FDA and others as development tools. Such tests are established and overseen similarly to other development tools such as biomarkers.

**Liability Under State Tort Law**

State tort law also provides a source of risk and of regulatory pressure for the developers and users of clinical AI systems, as well as other AI systems that could cause injury but that are not the focus of this section. Briefly, state tort law may make the developers or users of clinical AI systems liable when patients are injured as a result of using those systems. Such liability could come in the form of malpractice liability—that is, potential lawsuits against health providers, hospitals or other health care systems, and AI system developers for performing below the standard of care (Froomkin et al., 2019). Developers could also face product liability for defects in the design or manufacturing of AI systems or for failure to adequately warn users of the risks of a particular AI system. By imposing liability for injuries caused by AI systems when those injuries could have reasonably been avoided, whether by more careful development or more careful use, tort law exerts pressure on developers.

How exactly tort law will deal with clinical AI systems remains uncertain, because court decisions are retrospective and the technology is nascent. Tort law is principally grounded in state law, and its contours are shaped by courts on a case-by-case basis. This area will continue to develop. Three factors influencing tort liability are of particular note: the interaction of FDA approval and tort liability, liability insurance, and the impact of transparency on tort liability. To be clear, this area of law is still very much developing, and this section only sketches some of the ways different aspects of health care AI systems may interact with the system of tort liability.

**Interaction of FDA Approval and Tort Liability**

Different regulatory pathways influence the availability of state tort lawsuits against AI developers and, indirectly, the ability of state tort law (and liability insurers reacting to that law) to create independent incentives for the safe and effective development of clinical AI systems. In general, states may not establish statutory requirements that are “different from, or in addition to” FDA requirements regulating devices (21 U.S.C. § 360k). The U.S. Supreme Court has also held that this preempts certain state tort lawsuits alleging negligent design or manufacturing. For devices, including clinical AI apps, that undergo a full
premarket approval, state tort lawsuits are generally preempted under the Supreme Court’s holding in *Riegel v. Medtronic*, 552 U.S. 312 (2008). Nevertheless, this preemption will not apply to most AI apps, which are likely to be cleared through the 510(k) clearance pathway rather than premarket approval. Clearance under the 510(k) pathway will generally not preempt state tort lawsuits under the reasoning of *Medtronic v. Lohr*, 518 U.S. 470 (1996), because rather than directly determining safety and efficacy, FDA finds the new app to be equivalent to an already approved product. It is unclear what preemptive effect De Novo classification will have on preempting state tort lawsuits, because the Supreme Court emphasized both the thoroughness of premarket review and its determination that the device is safe and effective, rather than equivalent to an approved predicate device.

State tort lawsuits alleging violations of industry-wide requirements, such as CGMP or other validation requirements, are a contestable source of state tort liability. Some courts have found that lawsuits alleging violations of state requirements that parallel industry-wide requirements are preempted by federal law and that such violations may only be addressed by FDA. Other courts disagree, and the matter is currently unsettled (Tarloff, 2011). In at least some jurisdictions, if app developers violate FDA-imposed requirements, courts may find parallel duties under state law and developers may be held liable. Nevertheless, if app developers comply with all FDA-imposed industry-wide requirements, states cannot impose additional requirements.

**Liability Insurance**

The possibility of liability creates another avenue for regulation through the intermediary of insurance. Developers, providers, and health systems are all likely to carry liability insurance to decrease the risk of a catastrophic tort judgment arising from potential injury. Liability insurers set rules and requirements regarding what information must be provided or what practices and procedures must be followed in order to issue a policy. Although insurers are often not considered regulators, they can exert substantial, if less visible, pressure that may shape the development and use of clinical AI systems (Ben-Shahar and Logue, 2012).

**Impact of Transparency on Tort Liability**

Transparency and opacity also interact with tort liability. Determining causation can already be difficult in medical tort litigation, because injuries may result from a string of different actions and it is not always obvious which action or combination of actions caused the injury. Opacity in clinical AI systems may further complicate the ability of injured patients, lawyers, or providers or health
systems to determine precisely what caused the injury. Explainable algorithms may make it easier to assess tort liability, as could transparency around data provenance, training and validation methods, and ongoing oversight. Perversely, this could create incentives for developers to avoid certain forms of transparency as a way to lessen the likelihood of downstream tort liability. On the other hand, courts—or legislatures—could mandate that due care, in either the development or use of clinical AI tools, requires some form of transparency. To take a hypothetical example, a court might one day hold that when a provider relies on an algorithmic diagnosis, that provider can only exercise due care by assessing how the algorithm was validated. Developers or other intermediaries would then need to provide sufficient information to allow that assessment.

**PRIVACY, INFORMATION, AND DATA**

Regulation regarding patient privacy and data sharing is also highly relevant to AI development, implementation, and use, whether clinical AI or AI used for other health care purposes (“health care AI”). The United States lacks a general data privacy regime, but HIPAA includes a Privacy Rule that limits the use and disclosure of protected health information (PHI)—essentially any individually identifiable medical information—by covered entities (i.e., almost all providers, health insurers, and health data clearinghouses) and their business associates where the business relationship involves PHI (45 C.F.R. § 160.103). Covered entities and business associates may only use or disclose information with patient authorization, if the entity receives a waiver from an institutional review board or privacy board, or for one of several exceptions (45 C.F.R. § 164.502). These listed exceptions include the use and disclosure of PHI for the purposes of payment, public health, law enforcement, or health care operations, including quality improvement efforts but not including research aimed at creating generalizable knowledge (45 C.F.R. § 164.501). For health systems that intend to use their own internal data to develop in-house AI tools (e.g., to predict readmission rates or the likelihood of complications among their own patients), the quality improvement exception will likely apply. Even when the use or disclosure of information is permitted under HIPAA, the Privacy Rule requires that covered entities take reasonable steps to limit the use or disclosure to the minimum necessary to accomplish the intended purpose. While HIPAA does create protections for patient data, its reach is limited, and health information can come from many sources that HIPAA does not regulate (Price and Cohen, 2019).

A complex set of other laws may also create requirements to protect patient data. HIPAA sets a floor for data privacy, not a ceiling. State laws may be more restrictive; for instance, some states provide stronger protections for especially
sensitive information such as HIV status or substance abuse information (e.g., N.Y. Pub. Health Law § 2783). California’s Consumer Protection Act creates general protections for consumer information, including health data. And although the European Union’s GDPR focuses on actions that directly affect the European Union, it also places limits on the processing of data about EU residents, regardless of where the EU citizen resides globally, and may therefore affect the privacy practices of non-EU entities engaged in medical AI development (Marelli and Testa, 2018). The GDPR generally requires legal and real persons to collect and process only as much personal data as necessary, obtain such data only for a listed legitimate purpose or with consent, notify individuals of the receipt of data, and engage in privacy-centered policy design. Health data are especially protected under the GDPR, and their processing is prohibited unless with explicit consent or in a number of specified exceptions, such as for health operations or scientific research.

**Privacy and Patient Consent Issues in Health Care AI**

With regard to discrete clinical data, unstructured textual data, imagery data, waveform and time-series data, and hybrid data used in clinical AI models, the development and deployment of AI systems have complex interactions with privacy concerns and privacy law (e.g., Loukides et al., 2010). Adequate oversight of clinical AI systems must address the nature of potential privacy concerns wherever they may arise, approaches to address those concerns, and management of the potential tension between privacy and other governance concerns for clinical AI.

**Initial AI Development**

Privacy concerns occur in the first instance because training health care AI depends on assembling large collections of health data about patients (Horvitz and Mulligan, 2015). Health data about individuals are typically considered sensitive. Some forms of data are particularly sensitive, such as substance abuse data or sexually transmitted disease information (Ford and Price, 2016). Other forms of data raise privacy concerns about the particular individual, such as genetic data that can reveal information about family members (Ram et al., 2018). Collecting, using, and sharing patient health data raise concerns about the privacy of the affected individuals, whether those concerns are consequentialist (e.g., the possibility of future discrimination based on health status) or not (e.g., dignitary concerns about others knowing embarrassing or personal facts) (Price and Cohen, 2019). The process of collecting and sharing may also make data more vulnerable to interception or inadvertent access by other parties.
External Validation

External validation of clinical AI systems creates other avenues for privacy harms. Some proposals have called for third-party validation of medical AI recommendations and predictions to validate algorithmic quality (Ford and Price, 2016; Price, 2017a). Such an approach would either require making patient data available to those third parties or require the AI developer to have a partnership with a data owner, where data scientists ensure comparable data transformation and algorithm execution to provide external validation without direct data sharing.

Inference Generation

A third form of potential privacy harm that could arise from health care AI is quite different and involves the generation of inferences about individual patients based on their health data. Machine learning makes predictions based on data, and those predictions may themselves be sensitive data, or may at least be viewed that way by patients. In one highly publicized example of such a case, Target identified a teenage woman’s pregnancy based on changes in her purchasing habits and then sent her targeted coupons and advertisements, which led to her father learning of her pregnancy from Target before his daughter had shared the news (Duhigg, 2012). The epistemic status of this information is debatable; arguments have been made that inferences cannot themselves be privacy violations, although popular perception may differ (Skopek, 2018).

Some standard privacy-protecting approaches of data collectors and users face difficulties when applied to health care AI. The most privacy-protective approach limits initial data collection to necessarily limit the potential for problematic use or disclosure (Terry, 2017). However, this approach presumes that the data collector knows which data are necessary and which are not, knowledge that is often absent for health care AI.

De-Identification

De-identification, a common privacy-protecting approach, raises several concerns. Under the HIPAA Privacy Rule, patient information is not considered PHI (and thus not subject to the rule’s restrictions on use and disclosure) if a set of 17 listed pieces of identifying information have been removed (45 C.F.R. § 164.514(b)(2)(i)). These listed pieces include at least some elements that may be useful to health care AI, such as key dates, zip codes, or photographs of the patient. Thus, de-identification can lead to the loss of relevant data.

De-identification also raises two diametrically opposed concerns about gathering data. On the one hand, de-identification can lead to data fragmentation.
Patient data are gathered in many different contexts, including by different providers and different insurers. This diffuse data collection occurs both laterally, as patients encounter different parts of the medical system at the same time, and longitudinally, as patients shift between different medical environments over the course of time. Identifying information provides the easiest way to reassemble different parts of patient records into more comprehensive datasets that can help drive at least some forms of health care AI (e.g., long-term predictions of efficacy or mortality). When identifying information is removed from patient data, reassembly becomes harder, especially for developers with fewer resources. On the other hand, de-identification is not proof against re-identification (Ohm, 2010). Re-identification can happen at the level of the individual (via targeted efforts) or more broadly across datasets. “Data triangulation” refers to the idea that if data gatherers can collect multiple datasets that include some overlapping information, and if some of those datasets include identifying information, then data users can merge those datasets and identify individuals in the otherwise de-identified datasets (Mello and Cohen, 2018; Terry, 2017). Under current law, covered entities are limited in how they can re-identify data, since once it is re-identified it is again governed by HIPAA restrictions, but this does not govern those that are not covered entities. In addition, data-sharing agreements often include provisions prohibiting efforts at re-identification by the data recipient (Ohmann et al., 2017).

Individual consent and authorization provide the clearest possible path toward ameliorating privacy concerns but raise their own complications. When individuals know the purposes for which their information will be used and can give meaningful informed consent to those uses, privacy concerns can be limited. For machine learning and health care AI, however, future uses may be unpredictable. The revised Common Rule does allow for the provision of broad consent for unspecified future use (45 C.F.R. § 46.116). Nevertheless, systematic differences between those willing to consent to future data use and those unwilling to consent—or unable to consent because they lack that entry into the health data system—means that relying on individual authorization can introduce bias into datasets (Spector-Bagdady, 2016). Furthermore, the more meaningful the individual opportunity to consent, the higher the procedural hurdles created for the assembly of data—and the greater the likelihood of eventual bias. The Office of the National Coordinator for Health Information Technology has developed a Model Privacy Notice to “help developers convey information about their privacy and security policies” (ONC, 2018).

Data Infrastructure

The availability of data is an underlying legal and regulatory challenge for clinical AI system development, which requires large amounts of data for training
and validation purposes. Once particular AI systems are deployed in the real world, RWD should be collected to ensure that the AI systems are performing well and, ideally, to improve that performance. However, numerous hurdles exist to the collection of sufficient data (Price, 2016). Various privacy laws, as described above, restrict the collection of identifiable information, and de-identified information can be difficult to assemble to capture either long-term effects or data across different data sources. Informed consent laws, such as the Common Rule for federally funded research or the consent requirements incorporated into the GDPR, create additional barriers to data collection. Even where privacy or informed consent rules do not actually prohibit the collection, use, or sharing of data, some health care actors may limit such actions out of an abundance of caution, creating a penumbra of data limitations. In addition, for those actors who do find ways around these requirements, criticism and outrage may arise if patients feel they are inadequately compensated for their valuable data. On an economic level, holders of data have strong incentives to keep data in proprietary siloes to derive competitive advantage, leading to more fragmentation of data from different sources. For data holders who wish to keep data proprietary for economic reasons, referencing privacy concerns can provide a publicly acceptable reason for these tactics.

At least four possibilities emerge for collection of data, with some evidence of each in current practice:

1. **Large individual data holders:** Some large holders of individual data possess enough data to train AI models on their own, such as health systems (e.g., Partners or Ascension), health care payers (e.g., United Healthcare or Humana), or tech/data companies (e.g., Google or Apple).

2. **Data brokers and collaboration:** The collaboration or collection of data from different sources is possible, but these endeavors often encounter the hurdles described above, which may introduce limitations on data sources or bias in the incorporation process.

3. **Failure to collect data:** In some instances, no actor may have the incentive or ability to collect and gather data. This may be a problem especially for otherwise underserved populations, whose data may be under-represented in AI development and monitoring efforts.

4. **Government data infrastructure:** Governmental agencies can collect health data as part of an effort to support future innovation in clinical AI (among other efforts). The Precision Medicine Initiative’s All of Us cohort is an example of such an effort (Frey et al., 2016; NIH, 2019), as is the U.S. Department of Veterans Affairs’ Million Veteran Program (Gaziano et al., 2016), although the latter has more restrictive policies for its data use (Price, 2017b).
Of the four models, the first three are the straightforward results of current market dynamics. Each creates challenges, including smaller dataset size, potential bias in collection, access for other developers or for validators, and, in the case of failures to collect data, exclusion of some populations from AI development and validation. Government data infrastructure—that is, data gathered via government efforts for the purposes of fostering innovation, including clinical AI—has the greatest possibility of being representative and available for a variety of downstream AI uses but also faces potential challenges in public will for its collection. Even when the government itself does not collect data, it can usefully promulgate standards for data collection and consolidation (Richesson and Krischer, 2007; Richesson and Nadkarni, 2011); the lack of standards for EHRs, for instance, has led to persistent problems aggregating data across contexts.

**Tension Between Privacy and Data Access**

In general, there is tension between privacy-protecting approaches and access to big data for the development, validation, and oversight of health care AI. For instance, Google was sued for privacy violations in 2019 as a result of an agreement with the University of Chicago Medical Center to use the system’s data in AI and other big data applications (Cohen and Mello, 2019). Higher protections for patient data, whether regarding front-end collection or back-end use, increase the hurdles for the development of health care AI (Ford and Price, 2016). These hurdles may also exacerbate differences in capabilities between large, sophisticated entities—that is, health systems, health insurers, or large technology companies—and smaller developers that may lack the resources to develop AI in a privacy-protective fashion. However, privacy and innovation in health care AI are not in strict opposition. Newer technological approaches such as differential privacy (Malin et al., 2013) and dynamic consent (Kaye et al., 2014) can help enable development while still protecting privacy. In fact, the desire to protect privacy can be its own spur to the development of innovative technologies to collect, manage, and use health data. Nevertheless, resolving this tension presents a substantial ongoing challenge, one familiar in the development of a learning health system more generally. This resolution will not be simple and is beyond the scope of this chapter; it will demand careful policy making and continued engagement by stakeholders at various levels.

**KEY CONSIDERATIONS**

In summary, clinical AI tools present opportunities for improving patients’ and clinicians’ point-of-care decision making, and a viable business model is necessary to ensure that safe, effective clinical AI systems are developed, validated,
and sustainably deployed, implemented in EHR systems, and curated over time to maintain adequate accuracy and reliability. However, clinical AI systems could potentially pose risks in terms of inappropriate treatment recommendations, privacy breaches, or other harms (Evans and Whicher, 2018), and some types of clinical AI systems will be classified by regulatory agencies as SaMDs, subject to premarket clearance or approval and other requirements that aim to protect the public’s health. Other clinical AI tools may be deemed to be LDT-type services, subject to CLIA and similar regulations. Whatever agency is involved in oversight, compliance with regulations should be mandatory rather than voluntary, given the potential for problematic incentives for system developers (Evans and Whicher, 2018). As the law and policy of health care AI systems develop over time, it is both expected and essential that multiple stakeholders—including payers, patients and families, policy makers, diagnostic manufacturers and providers, clinicians, academics, and others—remain involved in helping determine how best to ensure that such systems advance the quintuple aim and improve the health care system more generally.

- The black box nature of a clinical AI system should not disqualify a system from regulatory approval or use, but transparency, where possible, can aid in oversight and adoption and should be encouraged or potentially required. AI systems, including black box systems, should be capable of providing the users with an opportunity to examine quantitative evidence that the recommendation in the current situation is indeed the best recent historical choice, supplying de-identified, aggregated data sufficient for the user to satisfy the user’s interest in confirming that this is so, or is at least no worse and no more uncertain than decisions the user would take independently were the AI not involved.

- When possible, machine learning–based predictive models should be evaluated in an independent dataset (i.e., external validation) before they are adopted in the clinical practice. Risk assessment to determine the degree to which dataset-specific biases affect the model should be undertaken. Regulatory agencies should recommend specific statistical methods for evaluating and mitigating bias.

- To the extent that machine learning–based models continuously learn from new data, regulators should adopt postmarket surveillance mechanisms to ensure continuing (and ideally improving) high-quality performance.

- Regulators should engage in collaborative governance efforts with other stakeholders and experts throughout the health system, including data scientists, clinicians, ethicists, and others, to continuously evaluate deployed clinical AI for effectiveness and safety on the basis of RWD.
• Government actors should invest in infrastructure that enables equitable, high-quality data collection, such as technical standards and technological capability building.
• Government actors should continue and increase efforts to develop large, high-quality, voluntary health datasets for clinical AI development (among other purposes), such as the All of Us cohort within the Precision Medicine Initiative, while ensuring adequate measures to address patient notice and potential harms.

REFERENCES


