GAO

Jointly published with

NATIONAL ACADEMY of MEDICINE

December 2019

United States Government Accountability Office

Science, Technology Assessment, and Analytics

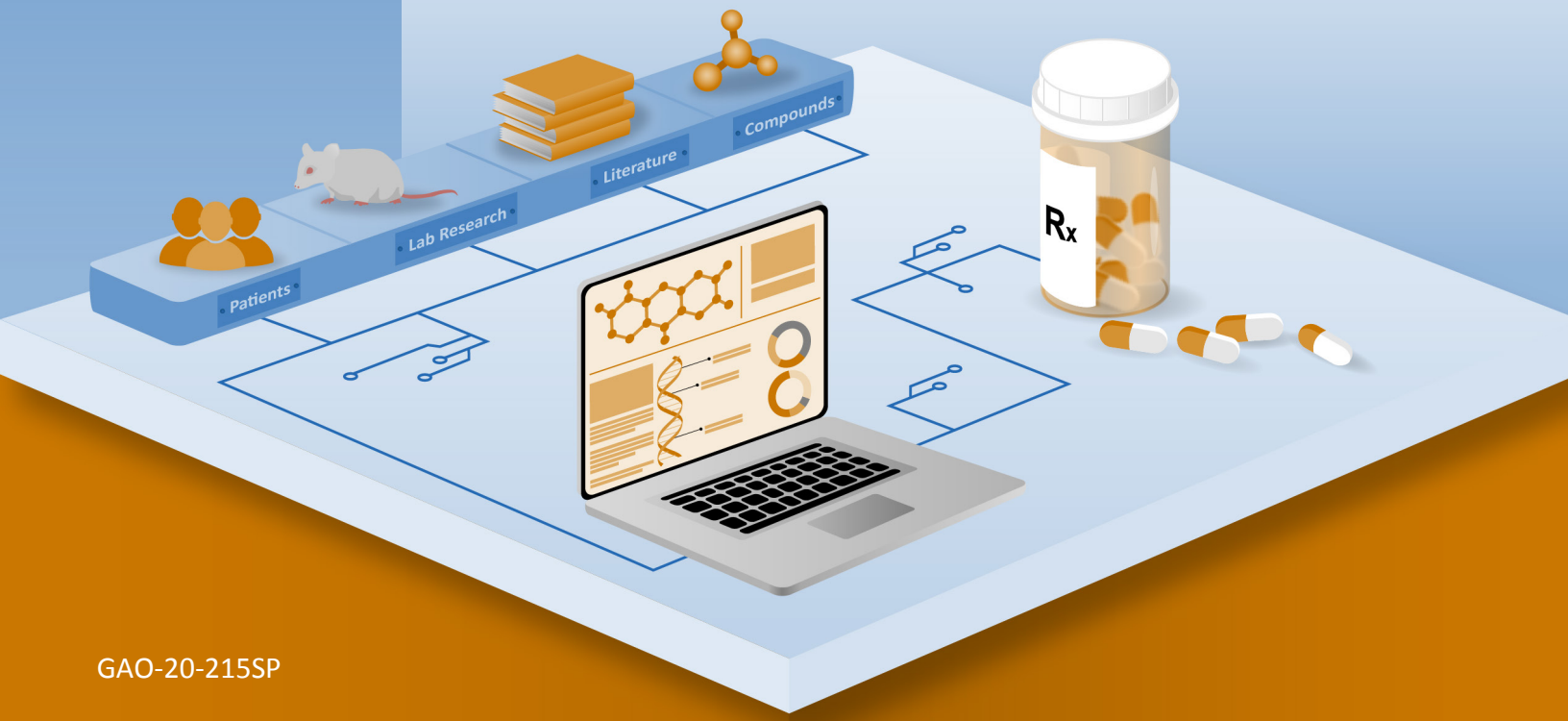
Report to Congressional Requesters

TECHNOLOGY ASSESSMENT

Artificial Intelligence in Health Care

Benefits and Challenges of Machine Learning in Drug Development

With field background content from the National Academy of Medicine



GAO-20-215SP

The cover image displays a stylized representation of data inputs from a variety of sources—including patient data from health records or clinical trials, lab research data, textual data from scientific and medical literature, and data on compounds and their properties—to a computer representing machine learning algorithms. Those algorithms then assist with various aspects of the drug development process, eventually resulting in a marketed drug.

This report is being jointly published by the Government Accountability Office (GAO) and the National Academy of Medicine (NAM). Part One of this joint publication presents material excerpted and adapted by NAM from its 2020 NAM Special Publication *Artificial Intelligence in Health Care: The Hope, the Hype, the Promise, the Peril*. Part Two is the full presentation of GAO's Technology Assessment *Artificial Intelligence in Health Care: Benefits and Challenges of Machine Learning in Drug Development*. Although GAO and NAM staff consulted with and assisted each other throughout this work, reviews were conducted by NAM and GAO separately and independently, and authorship of the text of Part One and Part Two of the report lies solely with NAM and GAO, respectively.

Foreword

The U.S. health care system is at an important crossroads as it faces major demographic shifts, burgeoning costs, and transformative technologies. Although the growth in health care costs has moderated recently, total annual health care spending in the United States is projected to reach nearly $6 trillion by 2027. Federal spending for health care programs—which accounts for more than a quarter of all health care spending—has grown faster than the overall economy in recent years, a trend projected to continue. Every day more than 10,000 Americans turn age 65, becoming eligible for Medicare. These demographic realities help illustrate the critical need to better address the effectiveness and efficiency of our nation's health care delivery systems.

Artificial intelligence and machine learning (AI/ML) is a set of technologies that includes automated systems able to perform tasks that normally require human intelligence, such as visual perception, speech recognition, and decision-making. AI/ML has promising applications in health care, including drug development. For example, it may have the potential to help identify new treatments, reduce failure rates in clinical trials, and generally result in a more efficient and effective drug development process. However, applying AI/ML technologies within the health care system also raises ethical, legal, economic, and social questions.

The Government Accountability Office (GAO) and the National Academy of Medicine (NAM), individually and in collaboration, have taken up the charge to explore AI/ML in health care, assess its implications, and identify key options available for optimizing its use. In recognition of mutual interests and obligations, and to reinforce and complement each other's work, NAM and GAO have cooperated on the development of two publications. The first is NAM's Special Publication: *Artificial Intelligence in Health Care: The Hope, the Hype, the Promise, the Peril*, adapted excerpts of which are presented as Part One of this joint publication. Any recommendations in Part One are those of NAM alone. The second is GAO's Technology Assessment: *Artificial Intelligence in Health Care: Benefits and Challenges of Machine Learning in Drug Development,* presented as Part Two.

This cooperative effort included two expert meetings, bringing diverse, interdisciplinary, and cross-sectoral perspectives to the discussions. We are grateful to the exceptionally talented staff of NAM and GAO as well as the experts, all of whom worked hard with enthusiasm, great skill, flexibility, clarity, and drive to make this joint publication possible.

Sincerely,

Timothy M. Persons, PhD
Chief Scientist and Managing Director,
Science, Technology Assessment, and Analytics
U.S. Government Accountability Office

J. Michael McGinnis, MD, MA, MPP
Leonard D. Schaeffer Executive Officer, and
Executive Director, NAM Leadership Consortium

GAO U.S. GOVERNMENT ACCOUNTABILITY OFFICE

NATIONAL ACADEMY OF MEDICINE

Executive Summary

This report is being jointly published by the Government Accountability Office (GAO) and the National Academy of Medicine (NAM). Part One of this joint publication presents material excerpted and adapted by NAM from its 2020 Special Publication *Artificial Intelligence in Health Care: The Hope, the Hype, the Promise, the Peril*. Part Two is the full presentation of GAO's Technology Assessment *Artificial Intelligence in Health Care: Benefits and Challenges of Machine Learning in Drug Development*. Although GAO and NAM staff consulted with and assisted each other throughout this work, reviews were conducted by NAM and GAO separately and independently, and authorship of the text of Part One and Part Two of this Executive Summary and the following report lies solely with NAM and GAO, respectively.

OVERVIEW OF PART ONE – NAM Special Publication: *Artificial Intelligence in Health Care: The Hope, the Hype, the Promise, the Peril*

The National Academy of Medicine's Special Publication: *Artificial Intelligence in Health Care: The Hope, the Hype, the Promise, the Peril* surveys current knowledge to present an accessible guide for relevant health care stakeholders such as artificial intelligence, machine learning (AI/ML) model developers, clinical implementers, clinicians, patients, and regulation and policy makers.[1] In this publication, an NAM expert working group comprised of leaders from various disciplines—public health, informatics, biomedical ethics, and implementation science—provides a sampling of present-day AI applications with a look to near-term possibilities, highlights the associated challenges and limitations, and outlines fundamental ethical, legal, regulatory, and societal considerations for the successful development and implementation of health care AI.

A key component shaping the publication was a January 2019 NAM convening of more than 60 experts from a range of stakeholder communities to consider how the draft could best ensure coverage of the most significant issues facing the development, deployment, or use of AI/ML in health care; that the

[1]Matheny, M., S. Thadaney, M. Ahmed, and D. Whicher, editors. *Artificial Intelligence in Health Care: The Hope, the Hype, the Promise, the Peril*. Washington, DC: National Academy of Medicine. Part One of this Joint Publication presents material excerpted and adapted by NAM from its 2020 Special Publication *Artificial Intelligence in Health Care: The Hope, the Hype, the Promise, the Peril*. Although GAO staff and leadership were consulted throughout the development process, authorship for the text lies solely with the National Academy of Medicine, the editors, and the authors (identified in the relevant sections and at the end of Part One).

solutions and approaches described and reviewed provided fair and balanced guidance for those interested in developing and deploying AI/ML models in health care settings; and the ways the content of the publication could most facilitate progress in the field. As an active participant, the GAO provided critical feedback on the content of the publication focusing on these dimensions.

Drawing on those discussions, and supplemented with written comments from external experts, the NAM publication identified several cross-cutting themes.

Potential Importance of AI/ML to Progress in Health and Health Care

With much of health and health care moving onto digital platforms, there has been a stunning growth in the volume of information generated through routine health-related processes and from products of health, health care, and biomedical science research. Especially as insights continue to emerge from exploration of underlying genetic predispositions to health and disease, the ability to use of AI and ML tools will soon be essential to assist with the growing field of precision medicine.

Furthermore, the ability to glean insights from the enormous body of data points generated daily from mobile apps (m-Health) and sensors will require the capacity for simultaneous data processing from multiple sources. The increasing availability of environmental and geospatial sensors developed on digital platforms contribute yet additional data universes requiring AI/ML before incorporation into predictive modeling tools.

AI and Transparency

As AI applications grow in their ability to lend perspective to health and health care decision-making, there is a compelling need for transparency in algorithms and data sources with the recognition that the need for algorithmic transparency is context-dependent, based on risk and intended use. For example, a high impact AI tool with immediate clinical implications warrants more stringent explanation requirements than a tool with a proven record of accuracy that is low risk and clearly conveys its recommendations to the end user. “Therefore, AI developers, implementers, users, and regulators should collaboratively define guidelines for clarifying the level of transparency needed across a spectrum.”[2]

[2]Matheny, Thadaney, Ahmed, and Whicher. *Artificial Intelligence in Health Care*.

As the field advances rapidly, regulators and legislators are required to remain nimble as they balance the complex interplay among AI innovation, safety, and trust. To avoid stymying AI development while ensuring proper oversight, regulators must engage myriad stakeholders and experts in the evaluation of clinical AI based on real-world data. As a harbinger of things to come, U.S. Food and Drug Administration (FDA) recently issued a framework for evaluating health care AI based on the level of patient risk, AI autonomy, and the dynamism of the tool.[3] Yet, to the extent that machine-learning models evolve with new data, issues of liability will continue to unfold with increasing involvement by the courts, regulators, and insurers.

Mitigating the hype

As the communication on the potential wonders of AI pervades social consciousness, it is easy for misguided fears and optimism to obscure its legitimate near-term possibilities. Although AI is certainly limited in its capacity to match the problem solving capacity of humans, AI-enabled automation is poised for disruptive workplace innovations. Given the necessary reliance on information technology (IT) and ML to help health professionals keep pace with the rapidly growing knowledge base, medical education will need a substantial overhaul. This needs to happen with an added focus on the use of AI as a routine decision-assistance tool. Training programs across multiple professions will require a focus on data science and the appropriate use of AI products and services. The bridging function to patient and consumer comfort levels with these emerging technologies will also need to be established to secure the bond of confidence between clinicians and their patients. Ultimately, the goal is to build competency in AI and data science to the point that health care AI provides an assistive benefit to humans rather than replacing them. For this reason, the near-term focus might be better termed "augmented intelligence."

Prioritizing Equity and Inclusivity

Among the many considerations in the NAM publication, especially strong emphasis was placed on the "the appropriate and equitable development and implementation of health care AI."[1] Prioritizing equity and inclusion begins with algorithms that have been developed from rich, population-representative datasets. Despite an abundance of health data, the lack of system interoperability and suboptimal data

[3]Food and Drug Administration. 2019. Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML) – Based Software as a Medical Device (SaMD). Available from: https://www.fda.gov/media/122535/download.

standardization techniques prevent the effective integration of health data from disparate systems. Without a robust base of data, AI algorithms fail to achieve successful levels of generalizability and utility.

Ensuring that equity, and inclusivity remain at the forefront in the deployment of AI requires the active engagement of system leaders, AI implementers, and regulators as they work to determine whether an AI tool is suitable for a particular environment and question whether its introduction could exacerbate existing biases and inequities. To address patient and community needs, health delivery organizations are in the process of developing (IT) governance strategies that expand linkages to social determinants and psychosocial data. National-scale efforts are needed to lower the barrier for adoption of these technologies and minimize the possible creation of a digital divide in underserved communities where IT capacities are less developed.

OVERVIEW OF PART TWO – GAO Technology Assessment: *Artificial Intelligence in Health Care: Benefits and Challenges of Machine Learning in Drug Development*

The GAO report *Artificial Intelligence in Health Care: Benefits and Challenges of Machine Learning in Drug Development* is the first in a planned series of technology assessments on the use of AI technologies in health care that GAO is conducting at the request of Congress.[4] This report discusses three topics: (1) current and emerging AI technologies available for drug development and their potential benefits, (2) challenges to the development and adoption of these technologies, and (3) policy options to address challenges to the use of machine learning in drug development. As one component of this review, NAM facilitated consultation with colleagues from the National Academies, to work closely with GAO in organizing a July 2019 meeting of 19 experts to explore these topics. NAM staff provided expertise, based on their work on the NAM Special Publication *Artificial Intelligence in Health Care: The Hope, the Hype, the Promise, the Peril*, to GAO in the identification of experts from federal agencies, academia, biopharmaceutical companies, machine learning-focused companies, and legal scholars. The meeting was intended to enhance GAO's understanding of ML in health care and drug development.

[4]Part Two of this Joint Publication presents the GAO Technology Assessment: *Artificial Intelligence in Health Care: Benefits and Challenges of Machine Learning in Drug Development.* Although NAM staff and leadership provided assistance and advice in the identification of issues and experts consulted during the development process (identified in app. II), responsibility for the text, findings, and options lies solely with GAO.

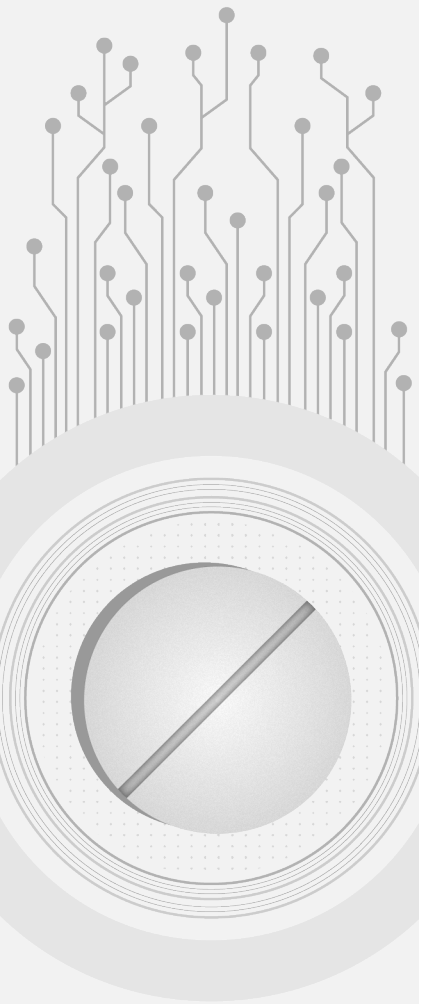
One of the report's high-level findings is that machine learning holds tremendous potential in drug development, according to stakeholders from government, industry, and academia. The current drug development process is lengthy and expensive, and can take 10 to 15 years to develop a new drug and bring it to market. ML techniques are already used throughout the drug development process and have the potential to expedite the discovery, design, and testing of drug candidates, decreasing the time and cost required. These improvements could save lives and reduce suffering by getting drugs to patients in need more quickly.

The technology assessment demonstrates the breadth of machine learning research and applications with examples from the first three steps of the drug development process—drug discovery, preclinical research, and clinical trials. In drug discovery, researchers are using ML to identify new drug targets, screen known compounds for new therapeutic applications, and design new drug candidates, among other applications. In preclinical research, ML can augment preclinical testing of drug candidates and predict toxicity before human testing. Researchers are also beginning to use ML to improve clinical trial design, a point where many drug candidates fail. These efforts include applying ML to patient selection and recruitment, and to identify patient populations who may react better to certain drugs, thus advancing towards the promise of precision medicine.

The technology assessment also identifies challenges that hinder the adoption and impact of machine learning in drug development, according to stakeholders, experts, and the literature. Gaps in research in biology, chemistry, and ML limit the understanding of and impact in this area. A shortage of high-quality data, which are required for ML to be effective, is another challenge. It is also difficult to access and share these data because of costs, legal issues, and a lack of incentives for sharing. Furthermore, a low supply of skilled and interdisciplinary workers creates hiring and retention challenges for drug companies. Lastly, uncertainty about regulation of machine learning used in drug development may limit investment in this field. Some of these challenges are similar to those identified in the NAM special publication, such as the lack of high-quality, structured data, and others are unique to drug development.

GAO describes options for policymakers—which GAO defines broadly to include federal agencies, state and local governments, academic and research institutions, and industry, among others—to use in addressing these challenges. In addition to the status quo, GAO identifies five policy options centered around research, data access, standardization, human capital, and regulatory certainty.

Table of Contents



PART ONE

Artificial Intelligence in Health Care:

Field Background

National Academy of Medicine (NAM)

Part One of this Joint Publication presents material excerpted and adapted by NAM from its 2020 Special Publication: *Artificial Intelligence in Health Care: The Hope, the Hype, the Promise, the Peril.* Although GAO staff and leadership were consulted throughout the development process, authorship of the text lies solely with the National Academy of Medicine, the editors, and the authors (identified in the relevant sections and at the end of Part One).

PART I: NAM FIELD OVERVIEW

Introduction: The emergence of artificial intelligence (AI) as a tool for better health care offers unprecedented opportunities to improve patient and clinical team outcomes, reduce costs, and impact population health. Many are already in use in health care. Nonetheless, the authors of the National Academy of Medicine's Special Publication titled *Artificial Intelligence in Health Care: The Hope, the Hype, the Promise, the Peril* not only underscore the promise, but also call out the issues for care and caution.

The material presented here has been adapted from the NAM's Special Publication and serves to provide a broad overview of current and near-term AI solutions; the challenges, limitations, and best practices for AI model development, adoption, and maintenance; the current legal and regulatory landscape for AI tools in health care; and prioritizes the need for equity, inclusion, and a human rights lens as we proceed together into a more technological future.

1. **Definitions of Key AI Terms:** The term **artificial intelligence** (AI), colloquially and in the scientific literature, takes on a range of meanings, from specific forms of AI, such as machine learning, to a hypothetical AI could be considered conscious or sentient. A formal definition of AI starts with the Oxford English Dictionary: "The capacity of computers or other machines to exhibit or simulate intelligent behavior; the field of study concerned with this." More nuanced definitions of AI might also consider what goal the AI is attempting to achieve and how it is pursuing that goal. In general, AI systems range from those that attempt to accurately model human reasoning to solve a problem, to those that ignore human reasoning and exclusively use large volumes of data to generate a framework to answer the question(s) of interest, to those that attempt to incorporate elements of human reasoning but do not require accurate modeling of human processes. The graphic below summarizes the domains of artificial intelligence *(Figure 1).*

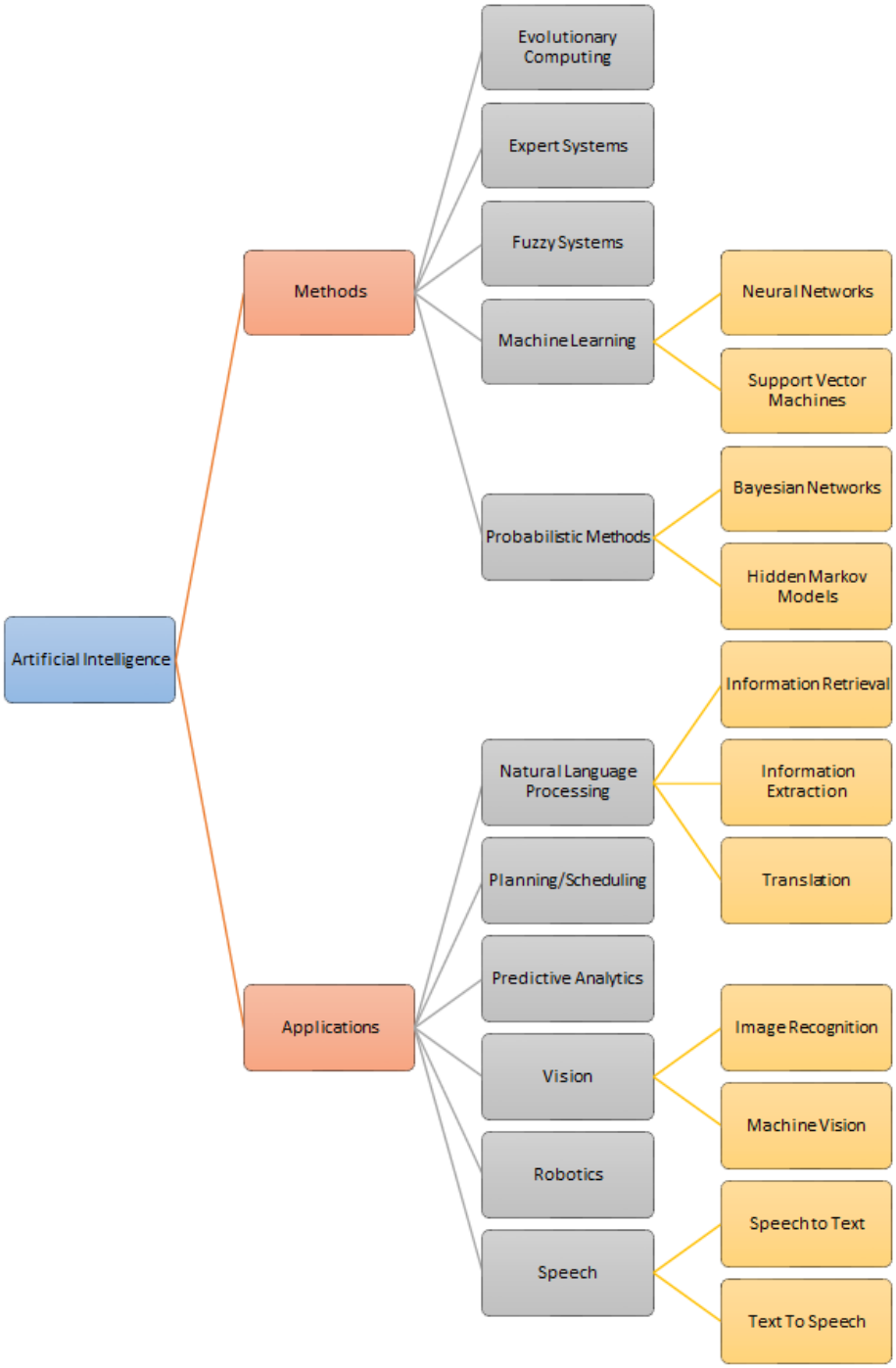


Figure 1 | A summary of the domains of artificial intelligence

SOURCE: Adapted with permission from a figure in Mills, M. 2015. Artificial Intelligence in Law—The State of Play in 2015? Legal IT Insider. https://www.legaltechnology.com/latest-news/artificial-intelligence-in-law-the-state-of-play-in-2015.

Machine learning is a family of statistical and mathematical modeling techniques that uses a variety of approaches to automatically learn and improve the prediction of a target state, without explicit programming (Witten et al., 2016). Different methods, such as Bayesian networks, random forests, deep learning, and artificial neural networks, each use different assumptions and mathematical frameworks for how data is ingested, and learning occurs within the algorithm. Regression analyses, such as linear and logistic regression, are also considered machine learning methods, although many users of these algorithms distinguish them from commonly defined machine learning methods (e.g., random forests, Bayesian Networks [BNs], etc.).

Natural language processing (NLP) enables computers to understand and organize human languages (Manning and Schütze, 1999). NLP needs to model human reasoning because it considers the meaning behind written and spoken language in a computable, interpretable, and accurate way. NLP incorporates rule-based and data-based learning systems, and many of the internal components of NLP systems are themselves machine learning algorithms with pre-defined inputs and outputs, sometimes operating under additional constraints. Examples of NLP applications include assessment of cancer disease progression and response to therapy among radiology reports (Kehl et al., 2019), and identification of post-operative complication from routine EHR documentation (Murff et al., 2011).

Expert systems are a set of computer algorithms that seek to emulate the decision-making capacity of human experts (Feigenbaum, 1992; Jackson, 1998; Leondes, 2002; Shortliffe and Buchanan, 1975). These systems rely largely on a complex set of Boolean and deterministic rules. An expert system is divided into a knowledge base, which encodes the domain logic, and an inference engine, which applies the knowledge base to data presented to the system to provide recommendations or deduce new facts.

Authors: Michael Matheny, MD, MS, MPH, Sonoo Thadaney Israni, MBA, Mahnoor Ahmed, MEng, and Danielle Whicher, PhD, MHS

2. **A Historical Perspective and Overview of Current AI:** If the term "artificial intelligence" might be given a birth date, it could be August 31, 1955, when John McCarthy, Marvin L. Minsky, Nathaniel Rochester, and Claude E. Shannon submitted "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence". The proposal and the resulting conference—the

1956 Dartmouth Summer Research Project on Artificial Intelligence—were the culmination of decades of thought by many others (Buchanan, 2005; Kline, 2011; Turing, 1950; Weiner, 1948). Although the conference produced neither formal collaborations nor tangible outputs, it helped galvanize the field (Moor, 2006).

Thought leaders in this era saw the future clearly, although optimism was substantially premature. In 1960, J. C. R. Licklider wrote, "The hope is that, in not too many years, human brains and computing machines will be coupled together very tightly, and that the resulting partnership will think as no human brain has ever thought and process data in a way not approached by the information-handling machines we know today" (Licklider, 1960).

By the 1970s, excitement gave way to disappointment because early successes that worked in well-structured, narrow problems failed to both generalize to broader problem solving and deliver operationally useful systems. The disillusionment, summarized in the ALPAC (Automatic Language Processing Advisory Committee) and Lighthill reports, resulted in an "AI Winter" with shuttered projects, evaporation of research funding, and general skepticism on the potential for AI systems (McCarthy, 1974; National Research Council, 1996).

Yet in health care, work continued. Iconic expert systems such as MYCIN (Shortliff, 1974) and others including Iliad, Quick Medical Reference, and Internist-1, were developed to assist with clinical diagnosis. AI flowered commercially in the 1980s, becoming a multibillion-dollar industry advising military and commercial interests (Miller, 1982; Sumner, 1993). However, all of these prospects ultimately failed to fulfill the hype and lofty promises, resulting in a second AI Winter from the late 1980s until the late 2000s

During this second AI Winter, the schools of computer science, probability, mathematics, and AI collaborated to overcome the initial failures of AI. In particular, techniques from probability and signal processing, such as Hidden Markov Models, Bayesian networks, and Stochastic search and optimization were incorporated into AI thinking, resulting in the field known as machine learning.

Around 2010, AI again regained prominence due to the success of machine learning and data science techniques, as well as significant increases in computational storage and power. These advances fueled the growth of technology titans such as Google and Amazon. Various ideas have laid the groundwork for artificial neural networks which have come to dominate the field of

machine learning. (Halevy et al., 2009; Krizhevsky, 2012). The resulting systems are called "deep learning systems" and show significant performance improvements over prior generations of algorithms for some use cases.

Modern AI has evolved from an interest in machines that think to ones that sense, think, and act. It is important to distinguish narrow from general AI. The popular conception of AI is of a computer, hypercapable in all domains, such as was seen even decades ago in science fiction with HAL 9000 in *2001: A Space Odyssey* (Stanley Kubrick, 1968) or aboard the USS *Enterprise* in the Star Trek franchise (Gene Roddenberry, 1966). These are examples of general AIs and, for now, are fictional. There is an active but niche general AI research community represented by Deepmind, Cyc, and OpenAI, among others. Narrow AI, in contrast, is an AI specialized at a single task, such as playing chess, driving a car, or operating a surgical robot.

Still, history has shown that AI has gone through multiple cycles of emphasis and disillusionment in use. It is critical that all stakeholders are aware and actively seek to educate and address public expectations and understanding of AI (and associated technologies) in order to manage hype and establish reasonable expectations, which will enable AI to be applied in effective ways that have reasonable opportunities for sustained success.

Authors: Jim Fackler, MD and Edmund Jackson, PhD

3. **How Artificial Intelligence Is Changing Health and Health Care:** The health care industry has been investing for years in technology solutions with the potential to transform health and health care There are promising examples, but there are gaps in the evaluation of these tools including AI, so it can be difficult to assess their impact. The NAM Special Publication reviews the potential of AI solutions for patients and families; the clinical care team; public health and population health program managers; business administrators; and researchers *(Figure 2).* Here we provide a sample of the potential solutions for patients and families, the clinical care team, and public health and population health program managers are detailed below.

Figure 2 | Examples of AI applications for stakeholder groups

Use Case/User Group	Category	Illustrative Examples of Applications	Technology
Patients and Families	• Health monitoring • Benefit/risk assessment	• Devices and wearables • Smartphone and tablet apps, websites	Machine learning, natural language processing (NLP), speech recognition, chatbots
	• Disease prevention and management	• Obesity reduction • Diabetes prevention and management • Emotional and mental health support	Conversational AI, NLP, speech recognition, chatbots
	• Medication management	• Medication adherence	Robotic home telehealth
	• Rehabilitation	• Stroke rehabilitation using apps and robots	Robotics
Clinical Care Teams	• Early detection, prediction, and diagnostics tools	• Imaging for cardiac arrhythmia detection, retinopathy • Early cancer detection (e.g., melanoma)	Machine Learning
	• Surgical Procedures	• Remote-controlled robotic surgery • AI-supported surgical roadmaps	Robotics, machine learning
	• Precision Medicine	• Personalized chemotherapy treatment	Supervised machine learning, reinforcement learning
	• Patient Safety	• Early detection of sepsis	Machine learning
Public Health Program Managers	• Identification of individuals at risk	• Suicide risk identification using social media	Deep learning (convolutional and recurrent neural networks)
	• Population health	• Eldercare monitoring	Ambient AI sensors
	• Population health	• Air pollution epidemiology • Water microbe detection	Deep learning, geospatial pattern mining, machine learning

Business Administrators	• International Classification of Diseases, 10th Rev. (ICD-10) coding	• Automatic coding of medical records for reimbursement	Machine learning, NLP
	• Fraud detection	• Health care billing fraud • Detection of unlicensed providers	Supervised, unsupervised, and hybrid machine learning
	• Cybersecurity	• Protection of personal health information	Machine learning, NLP
	• Physician management	• Assessment of physician competence	Machine learning, NLP
Researchers	• Genomics	• Analysis of tumor genomics	Integrated cognitive computing
	• Disease prediction	• Prediction of ovarian cancer	Neural networks
	• Discovery	• Drug discovery and design	Machine learning, computer-assisted synthesis

AI for Patients and Family: AI could soon play an important role in assisting patients and their families in the self-management of chronic diseases such as cardiovascular diseases, diabetes, and depression by assisting patients with taking medications, modifying diet, getting more physically active, assisting with care management, wound care, device management, and the delivery of injectables. Conversational agents, which can engage in two-way dialogue with the user via speech recognition, offer one example of how self-management of these diseases could be supplemented by AI solutions. Well known examples include Apple's Siri, Amazon's Alexa, or Microsoft's Cortana. Powered by NLP and natural language understanding, these interfaces may include text-based dialogue or present a human image (e.g., the image of nurse or coach) or a non-human image (e.g., a robot or animal) to provide a richer interactive experience. Conversational agents actually already exist to address depression, smoking cessation, asthma, and diabetes, although formal evaluation of these agents has been limited (Fitzpatrick et al., 2017).

In a more passive application for patients and families, AI can use raw data from accelerometers, gyroscopes, microphones, cameras, and smartphones for health monitoring and risk prediction. By using machine-learning algorithms to recognize patterns from the raw

data inputs and then categorize these patterns as indicators of an individual's behavior and health status, these systems can allow patients to understand and manage their own health and symptoms, as well as share data with medical providers. Consumer interest is high (~50%) in using data generated by apps, wearables, and Internet-of-Things devices to predict health risks (Accenture, 2018). Since 2013, AI start-up companies with a focus on health care and wearables have raised $4.3 billion to develop smart clothing, for example, bras designed for breast cancer risk prediction and other clothes for cardiac, lung, and movement sensing (Wiggers, 2018).

AI Solutions for the Clinical Care Team: There are two main areas of opportunity for AI in clinical care: (1) enhancing and optimizing care delivery and (2) improving information management, user experience, and cognitive support in EHRs. Prediction, early detection, and risk assessment for individuals is one of the most fruitful areas of AI applications (Sennaar, 2018). For example, diagnostic image recognition, which can be supported by AI applications, can differentiate between benign and malignant melanomas, diagnose retinopathy, identify cartilage lesions within the knee joint (Liu et al., 2018), detect lesion-specific ischemia, and predict node status after positive biopsy for breast cancer. Image recognition techniques can differentiate among competing diagnoses, assist in screening patients, and guide clinicians in radiotherapy and surgery planning (Matheson, 2018). AI platforms can, relatedly, provide roadmaps to assist surgical teams in the operating room, reducing risk and making surgery safer (Newmarker, 2018).

Clinicians are testing whether AI will permit them to personalize chemotherapy dosing and map patient response to a treatment to plan future dosing (Poon et al., 2018), a variation of precision medicine enabled by AI. AI-driven NLP has been used to identify polyp descriptions in pathology reports that trigger guideline-based clinical decision support to help clinicians determine the best surveillance intervals for colonoscopy exams (Imler et al., 2014). Other AI tools have helped clinicians select the best treatment options for complex diseases like cancer (Zauderer et al., 2014). Using retrospective data from other patients, AI techniques can predict treatment responses to different therapy combinations for an individual patient (Brown, 2018). These types of tools may serve to help select a treatment immediately, and may also provide new knowledge for future practice guidelines.

As genome-phenome integration is realized, the use of genetic data in AI systems for diagnosis, clinical care, and treatment planning will probably increase. To truly impact routine care,

though, genetic datasets will need to better represent the diversity of patient populations (Hindorff et al., 2018).

AI also has the potential to improve the way in which clinicians store and retrieve clinical documentation in EHRs. AI also has the potential to not only improve existing clinical decision support modalities, but to support improved cognitive support functions like smarter CDS alerts and reminders, as well as better access to peer-reviewed literature.

Population and Public Health Management: A spectrum of market-ready AI approaches to support population health programs already exists. They are used in areas of automated retinal screening, clinical decision support, predictive population risk stratification, and patient self-management tools (Contreras and Vehi, 2018; Dankwa-Mullan et al., 2018). Several solutions have received regulatory approval; for example, the U.S. Food and Drug Administration approved Medtronic's Guardian Connect, the first AI-powered continuous glucose monitoring system. Crowd-sourced, real-world data on inhaler use, combined with environmental data, led to a policy recommendation model that can be replicated to address many public health challenges by simultaneously guiding individual, clinical, and policy decisions. (Barrett et al., 2018) Other areas of potential overlap are standard risk prediction models that apply AI tools to facilitate recognition of clinically important but unanticipated predictor variables; and how AI can be used to not only predict risk, but also the presence or absence of a disease in an individual.

For public health professionals, the focus is on solutions for more efficient and effective administration of programs, policies, and services; disease outbreak detection and surveillance; as well as research. The range of AI solutions that can improve disease surveillance is considerable. For a number of years, researchers have tracked and refined the options for tracking disease outbreaks using search engine query data. Some of these approaches rely on the search terms that users type into internet search engines (e.g., Google Flu Trends, etc.).

At the same time, caution is warranted with these approaches. Relying on data not collected for scientific purposes (e.g., Internet search terms) to predict flu outbreaks has been fraught with error (Lazer et al., 2014). Non-transparent search algorithms that change constantly cannot be easily replicated and studied. These changes may occur due to business needs (rather than the needs of a flu outbreak detection application) or due to changes in the search behavior of

consumers. Finally, relying on such methods exclusively misses the opportunity to combine them and co-develop them in conjunction with more traditional methods. As Lazer et al. details, combining traditional and innovative methods (e.g., Google Flu Trends) performs better than either method alone.

AI and machine learning have also been used to develop a dashboard to provide live insight into opioid usage trends in Indiana (Bostic, 2018). This tool enabled prediction of drug positivity for small geographic areas (i.e., hot spots), allowing for interventions by public health officials, law enforcement, and program managers in targeted ways. A similar dashboarding approach supported by AI solutions has been used in Colorado to monitor HIV surveillance and outreach interventions and their impact after implementation (Snyder et al., 2016). This tool integrated data on regional resources with near real-time visualization of complex information to support program planning, patient management, and resource allocation.

Authors: Joachim Roski, PhD, MPH, Wendy Chapman, PhD, Jaimee Heffner, PhD, Ranak Trivedi, PhD, Guilherme Del Fiol, MD, PhD, Rita Kukafka, PhD, Paul Bleicher, MD, PhD, Hossein Estiri, PhD, Jeffrey Klann, PhD, and Joni Pierce, MBA, MS

4. **Potential Tradeoffs and Unintended Consequences of AI:** While we optimistically look to a future where AI-driven solutions can systematically improve health and medicine, AI systems could also have far-reaching unintended consequences and implications for patient populations, health systems, and the workforce. To mitigate the effect of these potential consequences, care must be given to the consideration of how tradeoffs between efficiency and equity impact populations in delivering against the unmet and unlimited demands of health care.

   The Future of Employment and Displacement: While anxiety over job losses due to AI and automation are likely exaggerated, advancing technology will almost certainly change roles as certain *tasks* are automated as seen in other industries. A conceivable future could see AI eliminating a clinician's need to perform manual tasks like checking patient vital signs (especially with self-monitoring devices), collecting laboratory specimens, preparing medications for pickup, transcribing clinical documentation, completing prior authorization forms, scheduling appointments, collecting standard history elements, and making routine diagnoses. However, most clinical jobs and patient needs require much more cognitive

adaptability, problem solving, and communication skills than a computer can muster. Despite the fear of AI eliminating jobs, industrialization and technology typically yield net productivity gains to society. For example, many assumed that automated teller machines (ATMs) would eliminate the need for bank tellers. Instead, the efficiencies gained through the use of ATMs enabled the expansion of bank branches and resulted in an even greater demand for tellers that could focus on higher cognitive tasks, such as interacting with customers, rather than simply counting money (Pethokoukis, 2016).

Need for Education and Workforce Development: A graceful transition into the AI era of health care that minimizes the unintended consequences of displacement will require deliberate redesigning of training programs. This ranges from support for a core basis of primary education in science, technology, engineering, and math literacy in the broader population to continuing professional education in the face of a changing environment. Health care workers in the AI future will need to learn how to use and interact with information systems, with foundational education in information retrieval and synthesis, statistics and evidence-based medicine appraisal, and interpretation of predictive models in terms of diagnostic performance measures. Institutional organizations (e.g., National Institutes of Health, health care systems, professional organizations, universities, and medical schools) should shift focus from skills that are easily replaced by AI automation to specific education and workforce development programs for work in the AI future, with emphasis in STEM, data science skills, and human skills that are hard to replace with technology.

AI System Augmentation of Human Tasks: While much of the popular discussion of AI focuses on how AI tools will replace human workers, realistically, in the foreseeable future, AI will function in an augmenting role, adding to the capabilities of the technology's human partners. As the volume of data and information available to inform patient care grows exponentially, AI tools will naturally become part of the clinical care team in much the same way a doctor is supported by a team of intelligent agents including specialists, nurses, physician assistants, pharmacists, social workers, and other health professionals (Meskó, Hetényi, and Győrffy, 2018). The technologies will be able to provide task-specific expertise in the data and information space, augmenting the capabilities of the physician and the entire team, making their jobs easier and more effective, and ultimately improving patient care (Herasevich, Pickering, and Gajic, 2018; Wu, 2019).

Hype versus Hope: One of the greatest near-term risks in the current development of AI tools in health care is not that it will cause serious unintended harm, but that it simply cannot meet the expectations stoked by excessive hype. Over the last decade, several factors have led to increasing interest and escalating hype of AI. Explicit advertising hyperbole may be one of the most direct triggers for unintended consequences of hype. While such promotion is important to drive interest and motivate progress, it can become counterproductive in excess. A combination of technical and subject domain expertise is needed to recognize the credible potential of AI systems and avoid the backlash that will come from overselling them.

Risks associated with model development and implementation: Since AI systems that will be deployed in the health care setting are constrained to learn from available observational health data, high fidelity and reliably measured outcomes are not always achievable. Although data from EHRs and other health information systems provide a rich longitudinal, multi-dimensional set of details about an individual's health, these data are often both noisy and biased as they are produced for different purposes in the process of documenting care. Poorly constructed or interpreted models from observational data can harm patients. Health care data scientists must be careful to apply the right types of modeling approaches based on the characteristics and limitations of the underlying data.

Although correlation can be sufficient for diagnosing problems and predicting outcomes in certain cases, methods that primarily learn associations between inputs and outputs can be unreliable, if not overtly dangerous when used to drive medical decisions. (Schulam and Saria, 2017) There are three common reasons why this is the case. First, performance of association-based models tend to be susceptible to even minor deviations between the development and implementation datasets. The learned associations may memorize dataset-specific patterns that do not generalize as the tool is moved to new environments where these patterns no longer hold. (Subbaswamy, Schulam, and Saria, 2019) A common example of this phenomenon is shifts in provider practice with the introduction of new medical evidence, technology, and epidemiology. If a tool heavily relies on a practice pattern to be predictive, as practice changes, the tool is no longer valid. (Schulam and Saria, 2017) Second, such algorithms cannot correct for biases due to feedback loops that are introduced when learning continuously over time. (Schulam and Saria, 2017) In particular, if the implementation of an AI system changes patient exposures, interventions, and outcomes (often as intended), it can cause data shifts that

degrade performance. Finally, the proposed predictors may be tempting to treat as factors one can manipulate to change outcomes but these are often misleading.

One approach is to update models over time so that they continuously adapt to local and recent data. Such adaptive algorithms offer constant vigilance and monitoring for changing behavior. However, this may exacerbate disparities when only well-resourced institutions can deploy the expertise to do so in an environment.

Training reliable models depends on training datasets being representative of the population where the model will be applied. Learning from real world data---where insights can be drawn from patients similar to a given index patient---has the benefit of leading to inferences that are more relevant, but it is important to characterize populations where there is inadequate data to support robust conclusions. For example, a tool may show acceptable performance on average across individuals captured within a data set, but may perform poorly for specific subpopulations because the algorithm has not had enough data to learn from. In genetic testing, minority groups can be disproportionately adversely affected when recommendations are made based on data that does not adequately represent them. (Manrai et al., 2016) Test-time auditing tools that can identify individuals for whom the model predictions are likely to be unreliable can reduce the likelihood of incorrect decision-making due to model bias. (Schulam and Saria, 2017)

Machine learning that relies on observational data could also generally have an amplifying effect on existing behavior, regardless of whether that behavior is beneficial or exacerbates existing societal biases. For instance, a study found that machine translation systems were biased against women due to the way in which women were described in the data used to train the system. (Prates, Avelar, and Lamb, 2018) While some of these algorithms were revised or discontinued, the underlying issues will continue to be significant problems, requiring constant vigilance, as well as algorithm surveillance and maintenance to detect and address.

AI Systems Transparency: Transparency is a key theme that underlies deeper issues related to privacy and consent or notification for patient data use, and to potential concerns on the part of patients and clinicians around being subject to algorithmically-driven decisions. Consistent progress in the development and adoption of AI in health care will only be feasible if health care consumers and health care systems are mutually recognized as trusted data partners.

Tensions exist among the desire for robust data aggregation to facilitate the development and validation of novel AI models, the need to protect consumer privacy, and the need to demonstrate respect for consumer preferences through informed consent or notification procedures. However, lack of transparency about data use and privacy practices could create a situation in which patients do not clearly consent to their data being used in ways they do not understand, realize, or accept. Current consent practices for the use of EHR and claims data are generally based on models focused on HIPAA privacy rules, and some argue that HIPAA needs updating (Mello and Cohen, 2018). The progressive integration of other sources of patient-related data (e.g., genetic information, social determinants of health), and the facilitated access to highly granular and multi-dimensional data are changing the protections provided by traditional mechanisms, such as HIPAA. For instance, with more data available, re-identification becomes easier to perform (Cohen and Mello, 2019). Regulations need to be updated and consent processes will need to be more informative of those added risks.

Authors: Jonathan Chen, MD, PhD, Andrew Beam, PhD, Suchi Saria, PhD, and Eneida Mendonca, MD, PhD

5. **Best Practices for Machine-Learning Model Development and Validation:** Machine learning models should be thoughtfully developed and validated. First, all stakeholders must understand the needs of clinical practice, so that proposed AI systems address the practicalities of health care delivery. Second, it is necessary that such models be developed and validated through a team effort, involving AI experts and health care providers. Throughout the design and validation process, it is important to be mindful of the fact that the datasets used to train AI are heterogeneous, complex, and nuanced in ways that are often subtle and institution-specific. This impacts how AI tools are monitored for safety and reliability, and how they are adapted for different locations and over time. Third, before deployment at the point of care, AI systems should be rigorously evaluated to ensure their competency and safety, in a similar process to that done for drugs, medical devices, and other medical interventions.

   Establishing Utility: When considering the use of AI in health care, it is necessary to know how a member of the care team would act, given a model's output. While model evaluation typically focuses on metrics, such as positive predictive value, sensitivity (or recall), specificity, and

calibration, constraints on the action triggered by the model's output (e.g. continuous rhythm monitoring might be constrained by availability of Holter monitors) often can have a much larger influence in determining model utility (Moons et al., 2012). Completing model selection, then doing a net-benefit analysis, and later factoring work constraints is suboptimal (Shah et al., 2019). Realizing the benefit of implementation of AI into the work flow requires defining potential utility upfront. Only by including the characteristics of actions taken on the basis of the model's predictions, and factoring in their implications, can a model's potential usefulness in improving care be properly assessed.

Learning a Model: After the potential utility of the model has been established, model developers and model users need to interact closely when learning a model because many modeling choices are dependent on the model's context of use (Wiens et al., 2019). For example, the need for external validity depends on what one wishes to do with the model, the degree of agency ascribed to the model, and the nature of the action triggered by the model.

It is well known that biased data will result in biased models. Thus, the data that is selected to learn from matters far more than the choice of the specific mathematical formulation of the model. Model builders need to pay close attention to the data they train on and to think beyond the technical evaluation of models. Even in technical evaluation, it is necessary to look beyond the ROC curves, and examine multiple dimensions of performance. For decision making in the clinic, additional metrics such as calibration, net reclassification, and a utility assessment are necessary. Given the nonobvious relationship between a model's positive predictive value, recall, and specificity to its utility, it is important to examine simple and obvious parallel baselines, such as a penalized regression model applied on the same data that are supplied to more sophisticated models such as deep learning.

The topic of interpretability deserves special discussion because of ongoing debates around interpretability, or the lack of it (Licitra, Trama, and Hosni, 2017; Lipton, 2016; Voosen, 2017). To the model builder, interpretability often means the ability to explain which variables and their combinations, in what manner, led to the output produced by the model (Friedler et al., 2019). To the clinical user, interpretability could mean one of two things: a sufficient enough understanding of what is going on, so that they can trust the output and/or be able to get liability insurance for its recommendations; or enough causality in the model structure to provide hints as to what mitigating action to take. To avoid wasted effort, it is important to

understand what kind of interpretability is needed in a particular application. A black box model may suffice if the output is trusted, and trust can be obtained by prospective assessment of how often the model's predictions are correct and calibrated.

Data Quality: Bad data quality adversely impacts patient care and outcomes (Jamal, McKenzie, and Clark, 2009). A recent systematic review shows that the AI models could dramatically improve if four particular adjustments were made: the use of multicenter datasets, incorporation of time varying data, assessment of missing data as well as informative censoring, and development of metrics of clinical utility (Goldstein et al., 2017). As a reasonable starting point for minimizing data quality issues, the authors of the NAM Special Publication recommend that data should adhere to the FAIR (findability, accessibility, interoperability, and reusability) principles in order to maximize the value of data (Wilkinson et al., 2016). An often-overlooked detail is when and where certain data become available and whether the mechanics of data availability and access are compatible with the model being constructed.

Stakeholder education and managing expectations: The use of AI solutions presents a wide range of challenges to law and ethics, most of which are still being worked out. For example, when a physician makes decisions assisted by AI, it is not always clear where to place blame in the case of failure. This subtlety is not new to recent technological advancements, and in fact was brought up decades ago (American Journal of Bioethics, 2010). However, most of the legal and ethical issues were never fully addressed in the history of computer-assisted decision support, and a new wave of more powerful AI-driven methods only adds to the complexity of ethical questions (e.g., the frequently condemned black box model) (Char et al., 2018).

Model builders need to better understand the datasets they choose to learn from. Decision makers need to look beyond technical evaluations and ask for utility assessments. Media needs to do a better job in articulating both immense potential and the risks of adopting the use of AI in health care. Therefore, it is important to promote a measured approach to adopting AI technology, which would further AI's role as augmenting rather than replacing human actors. This framework could allow the AI community to make progress while managing evaluation challenges (e.g., when and how to employ interpretable models versus black-box models) as well as ethical challenges that are bound to arise as the technology gets widely adopted.

Authors: Hongfang Liu, PhD, Hossein Estiri, PhD, Jenna Wiens, PhD, Anna Goldenberg, PhD, Suchi Saria, PhD, and Nigam Shah, MBBS, PhD

6. **Deploying AI in Clinical Settings:** For AI deployment in health care practice to be successful, it is critical that the lifecycle of AI use be overseen through effective governance. **IT governance is the set of** processes that ensure the effective and efficient use of IT in enabling an organization to achieve its goals by overseeing the evaluation, selection, prioritization, and funding, implementation, and tracking of IT projects. Another facet of IT governance that is relevant to AI is data governance, which institutes methodical process that an organization adopts to manage its data and ensure the data meet specific standards and business rules before entering them into a data management system. A health care enterprise that seeks to leverage AI should consider, characterize, and adequately resolve a number of key considerations prior to moving forward with the decision to develop and implement an AI solution *(see Figure 3)*.

Figure 3 | Key Considerations for Instructional Infrastructure and Governance

Consideration	Relevant Governance Questions
Organizational Capabilities	Does the organization possess the necessary technologic (e.g., IT infrastructure, IT personnel) and organizational (knowledgeable and engaged workforce, educational and training capabilities) to adopt, assess and maintain AI driven tools?
Data Environment	What data are available for AI development? Do current systems possess the adequate capacity for storage, retrieval, and transmission to support AI tools?
Interoperability	Does the organization support and maintain data at rest and in motion per national and local standards for interoperability (e.g., SMART on FHIR)?
Personnel Capacity	What expertise exists in the health care system to develop and maintain the AI algorithms?
Cost, Revenue, and Value	What will be the initial and ongoing costs to purchase, install, and train users, to maintain underlying data models, and to monitor for variance in model performance? Is there an anticipated return on investment from the AI deployment? What is the perceived value for the institution related to AI deployment?

Safety and Efficacy Surveillance	Are there governance and processes in place to provide regular assessments of the safety and efficacy of AI tools?
Patient/Family/Consumer Engagement	Does the institution have in place formal mechanisms for patient/family/consumer such a council or advisory board that can engage and voice concerns on relevant issues related to implementation, evaluation etc.?
Cybersecurity and Privacy	Does the digital infrastructure for health care data in the enterprise have sufficient protections in place to minimize the risk of breaches of privacy if AI is deployed?
Ethics and Fairness	Is there an infrastructure in place at the institution to provide oversight and review of AI tools to ensure that the known issues related to ethics and fairness are addressed and that vigilance for unknown issues is in place?
Regulatory Issues	Are there specific regulatory issues that must be addressed and if so, what type of monitoring and compliance programs will be necessary?

Organizational Approach to Implementation: AI development and implementation should follow established best practice frameworks in implementation science and software development. Frameworks for conceptualizing, designing and evaluating this process are discussed in more detail in the NAM Special Publication, but all implicitly incorporate the most fundamental basic health care improvement model, often referred to as a plan-do-study-act (PDSA) cycle first introduced by W.E. Deming more than two decades ago (Deming, 2000). The PDSA cycle relies on the intimate participation of employees involved in the work, detailed understanding of workflows, and careful ongoing assessment of implementation that informs iterative adjustments. Newer methods of quality improvement introduced since Deming represent variations or elaborations of this approach. All too often, however, quality improvement efforts frequently fail because they are focused narrowly on a given task or set of tasks using inadequate metrics without due consideration of the larger environment in which change is expected to occur (Muller, 2018).

Such concerns are certainly relevant to AI implementation. New technology promises to substantially alter how medical professionals currently deliver health care at a time when morale in the workforce is generally poor (Shanafelt et al., 2012). One of the challenges of the use of AI in health care is that integrating it within the EHR and improving existing decision and workflow support tools may be viewed as an extension of an already unpopular technology

(Sinsky et al., 2016). Moreover, there are a host of concerns that are unique to AI, some well and others poorly founded, which might add to the difficulty of implementing AI applications.

In recognition that basic quality improvement approaches are generally inadequate to produce large-scale change, the field of implementation science has arisen to characterize how organizations can undertake change in a systematic fashion that acknowledges their complexity. Some frameworks are specifically designed for evaluating the effectiveness of implementation, such as the Consolidated Framework for Implementation Research (CFIR) or the Promoting Action on Research Implementation in Health Services (PARiHS). In general, these governance and implementation frameworks emphasize sound change management and methods derived from implementation science that undoubtedly apply to implementation of AI tools (Damschroder et al., 2009; Rycroft-Malone, 2004).

Clinical Outcome Monitoring: The complexity and extent of local evaluation and monitoring may necessarily vary depending on the way AI tools are deployed into the clinical workflow, the clinical situation, and the type of CDS being delivered, as these will in turn define the clinical risk attributable to the AI tool.

For higher risk AI tools, a focus on clinical safety and effectiveness—from either a non-inferiority or superiority perspective—is of paramount importance even as other metrics (e.g., API data calls, user experience information) are considered. High-risk tools will likely require evidence from rigorous studies for regulatory purposes and will certainly require substantial monitoring at the time of and following implementation. For low-risk clinical AI tools used at point of care, or those that focus on administrative tasks, evaluation may rightly focus on process of care measures and metrics related to the AI's usage in practice to define its positive and negative effects. The authors of the NAM Special Publication strongly endorse implementing all AI tools using experimental methods (e.g., randomized controlled trials or A/B testing) where possible. Large-scale pragmatic trials at multiple sites will be critical for the field to grow but may be less necessary for local monitoring and for management of an AI formulary. In some instances, due to feasibility, costs, time constraints or other limitations, a randomized trial may not be practical or feasible. In these circumstances quasi-experimental approaches such as stepped-wedge designs or even carefully adjusted retrospective cohort studies, may provide valuable insights.

Monitoring outcomes after implementation will permit careful assessment, in the same manner that systems regularly examine drug usage or order sets and may be able to utilize data that are innately collected by the AI tool itself to provide a monitoring platform. Recent work has revealed that naive evaluation of AI system performance may be overly optimistic, providing a need for more thorough evaluation and validation.

Clinical AI performance can also deteriorate within a site when practices, patterns, or demographics change over time. As an example, consider the policy by which physicians order blood lactate measurements. Historically, it may have been the case that, at a particular hospital, lactate measurements were only ordered to confirm suspicion of sepsis. A clinical AI tool for predicting sepsis that was trained using historical data from this hospital would be vulnerable to learning that the act of ordering a lactate measurement is associated with sepsis rather than the elevated value of the lactate. However, if hospital policies change and lactate measurements are more commonly ordered, then the association that had been learned by the clinical AI would no longer be accurate. Alternatively, if the patient population shifts, for example to include more drug users, then elevated lactate might become more common and the value of lactate being measured would again be diminished. In both the case of changing policy or patient population, performance of the clinical AI application is likely to deteriorate, resulting in an increase of false positive sepsis alerts.

More broadly, such examples illustrate the importance of careful validation in evaluating the reliability of clinical AI. A key means for measuring reliability is through validation on multiple datasets. Classical algorithms that are applied natively or used for training AI are prone to learning artifacts specific to the site that produced the training data or specific to the training dataset itself. There are many subtle ways that site-specific or dataset-specific bias can occur in real world datasets. Validation using external datasets will show reduced performance for models that have learned patterns that do not generalize across sites (Schulam and Saria, 2017).

In addition to monitoring overall measures of performance, evaluating performance on key patient subgroups can further expose areas of model vulnerability: High average performance overall is not indicative of high performance across *every* relevant subpopulation. Careful examination of stratified performance can help expose subpopulations where the clinical AI model performs poorly and therefore poses higher risk. Further, tools that detect individual

points where the clinical AI is likely to be uncertain or unreliable can flag anomalous cases. By introducing a manual audit for these individual points, one can improve reliability during use (e.g., Soleimani, Hensman, and Saria, 2018 and Schulam and Saria, 2019). Traditionally, uncertainty assessment was limited to the use of specific classes of algorithms for model development. However, recent approaches have led to wrapper tools that can audit some black box models (Schulam and Saria, 2019). Logging cases flagged as anomalous or unreliable and performing a review of such cases from time to time may be another way to bolster post marketing surveillance, and FDA requirements for such surveillance could require such techniques.

Authors: Stephan D. Fihn, MD, MPH, Suchi Saria, PhD, Eneida Mendonca, MD, PhD, Seth Hain, MS, Michael Matheny, MD, MS, MPH, Nigam Shah, MBBS, PhD, Hongfang Liu, PhD, and Andrew Auerbach, MD

7. **Conclusion:** AI in health care is poised to make transformative and disruptive advances in health care. It is prudent to balance the need for thoughtful, inclusive health care AI that plans for and actively manages and reduces potential unintended consequences, while not yielding to marketing hype and profit motives. The straightforward path for AI is to start with real problems in health care, explore the best solutions by engaging relevant stakeholders, frontline users, patients and their families—including AI and non-AI options—and implement and scale the ones that meet a new Quintuple Aim of equity and inclusion *(See Figure 4).*

Figure 4 | Advancing the Quintuple Aim

SOURCE: Matheny, M., S. Thadaney, M. Ahmed, and D. Whicher, editors. *Artificial Intelligence and Health Care: The Hope, the Hype, the Promise, and the Perils.* Washington, DC: National Academy of Medicine.

In *21 Lessons for the 21st Century*, Yuval Noah Harari writes, "Humans were always far better at inventing tools than using them wisely" (Harari, 2018, p. 7). It is up to us, the stakeholders, experts, and users of these technologies, to ensure that they are used in an equitable and appropriate fashion to uphold the human values that inspired their creation—that is, better health and wellness for all.

BIBLIOGRAPHY

Accenture. 2018. *Consumer Survey on Digital Health: US Results.* https://www.accenture.com/t20180306T103559Z__w__/us-en/_acnmedia/PDF-71/accenture-health-2018-consumer-survey-digital-health.pdf (accessed November 12, 2019).

Barrett, M., V. Combs, J. G. Su, K. Henderson, M. Tuffli, and AIR Louisville Collaborative. 2018. AIR Louisville: Addressing asthma with technology, crowdsourcing, cross-sector collaboration, and policy. *Health Affairs (Millwood)* 37(4):525–534.

Berg, J. 2010. Review of "The Ethics of Consent, eds. Franklin G. Miller and Alan Wertheimer." *American Journal of Bioethics* 10(7):71–72.

Bostic, B. 2018. Using artificial intelligence to solve public health problems. *Beckers Hospital Review*. https://www.beckershospitalreview.com/healthcare-information- technology/using-artificial-intelligence-to-solve-public-health-problems.html (accessed November 12, 2019).

Brown, N., and T. Sandholm. 2018. Superhuman AI for heads-up no-limit poker: Libratus beats top professionals. *Science* 359:418–424.

Buchanan, B. G. 2005. A (very) brief history of artificial intelligence. *AI Magazine* 26(4):53–60.

Char, D.S., N. H. Shah, and D. Magnus, D., 2018. Implementing machine learning in health care—addressing ethical challenges. *New England Journal of Medicine*, 378(11), 981.

Cohen, G., and M. Mello. 2019. Big data, big tech, and protecting patient privacy. *JAMA* 322(12):1141–1142.

Contreras, I. and J. Vehi. 2018. Artificial intelligence for diabetes management and decision support: literature review. *Journal of medical Internet research*, 20(5), .e10775.

Damschroder, L. J., D. C. Aron, R. E. Keith, S. R. Kirsh, J. A. Alexander, and J. C. Lowery. 2009. Fostering implementation of health services research findings into practice: A consolidated framework for advancing implementation science. *Implementation Science* 4(1):50.

Dankwa-Mullan, I., M. Rivo, M. Sepulveda, Y. Park, J. Snowdon, and K. Rhee. 2019. Transforming diabetes care through artificial intelligence: the future is here. *Population health management*, 22(3), 229-242.

Deming, W. E. 2000. *The New Economics for Industry, Government, and Education. Cambridge,* MA: MIT Press. http://www.ihi.org/resources/Pages/Publications/NewEconomicsforIndustryGovernmentEducation.aspx (accessed November 13, 2019).

Feigenbaum, E. 1992. Expert Systems: Principles and Practice. *The Encyclopedia of Computer Science and Engineering*.

Fitzpatrick, K. K., A. Darcy, and M. Vierhile. 2017. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agenda (Woebot): A randomized controlled trial. *JMIR Mental Health* 4(2):e19.

Friedler, S. A., C. D. Roy, C. Scheidegger, and D. Slack. 2019. Assessing the local interpretability of machine learning models. *arXiv.org*. https://ui.adsabs.harvard.edu/abs/2019arXiv190203501S/abstract (accessed November 13, 2019).

Goldstein, B. A., A. M. Navar, M. J. Pencina, and J. Ioannidis. 2017. Opportunities and challenges in developing risk prediction models with electronic health records data: A systematic review. *Journal of the American Medical Informatics Association* 24(1):198–208.

Halevy, A., P. Norvig, and F. Pereira. 2009. The unreasonable effectiveness of data. *IEEE Intelligent Systems* 24:8–12. https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/35179.pdf (accessed November 8, 2019).

Harari, Y. N. 2018. *21 Lessons for the 21st Century*. Random House: New York, NY, USA.

Herasevich, V., B. Pickering, and O. Gajic. 2018. How Mayo Clinic is combating information overload in critical care units. *Harvard Business Review*. https://www.bizjournals.com/albany/news/2018/04/23/hbr-how-mayo-clinic-is-combating-information.html (accessed November 13, 2019).

Hindorff, L.A., Bonham, V.L., Brody, L.C., Ginoza, M.E., Hutter, C.M., Manolio, T.A. and Green, E.D., 2018. Prioritizing diversity in human genomics research. *Nature Reviews Genetics*, 19(3), 175.

Imler, T. D., J. Morea and T. F. Imperiale. 2014. Clinical decision support with natural language processing facilitates determination of colonoscopy surveillance intervals. *Clinical Gastroenterology and Hepatology*, 12(7), 1130-1136.

Jackson, P. 1998. *Introduction to Expert Systems*. Boston: Addison-Wesley Longman Publishing Co., Inc.

Jamal, A., K. McKenzie, and M. Clark. 2009. The impact of health information technology on the quality of medical and health care: A systematic review. *Health Information Management Journal* 38(3):26–37.

Kehl, K.L., H. Elmarakeby, M. Nishino, E. M. Van Allen, E.M. Lepisto, M. J. Hassett, B. E. Johnson, and D. Schrag. 2019. Assessment of deep natural language processing in ascertaining oncologic outcomes from radiology reports. *JAMA Oncology*, 5(10), 1421-1429.

Kline, R. R. 2011. Cybernetics, automata studies, and the Dartmouth Conference on Artificial Intelligence. *IEEE Annals of the History of Computing* 33(4):5–16.

Krizhevsky, A., I. Sutskever, and G. E. Hinton. 2012. ImageNet classification with deep convolutional neural networks. Presented at 25th International Conference on Neural Information Processing Systems. https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf (accessed November 8, 2019).

Lazer, D., R. Kennedy, G. King, G. and A. Vespignani. 2014. The parable of Google Flu: traps in big data analysis. *Science*, 343(6176), 1203-1205.

Lee, K. J. 2018. AI device for detecting diabetic retinopathy earns swift FDA approval. *American Academy of Ophthalmology ONE Network*. https://www.aao.org/headline/first-ai-screen-diabetic-retinopathy-approved-by-f (accessed November 13, 2019).

Leondes, C. T. 2002. *Expert Systems: The Technology of Knowledge Management and Decision Making for the 21st century*. San Diego: Academic Press.

Licitra, L., A. Trama, and H. Hosni. 2017. Benefits and risks of machine learning decision support systems. *JAMA* 318(23):2354–2354.

Licklider, J. C. R. 1960. Man-computer symbiosis. *IRE Transactions on Human Factors in Electronics*. HFE-1:4–11.

Lipton, Z. C. 2016. The mythos of model interpretability. *arXiv.org*. https://arxiv.org/abs/1606.03490 (accessed November 13, 2019).

Liu, F., Z. Zhou, A. Samsonov, D. Blankenbaker, W. Larison, A. Kanarek, K. Lian, S. Kambhampati, and R. Kijowski. 2018. Deep learning approach for evaluating knee MR images: achieving high diagnostic performance for cartilage lesion detection. *Radiology*, 289(1), 160-169.

Manning, C. D., and H. Schütze. 1999. *Foundations of statistical natural language processing*. MIT press.

Manrai, A. K., B. H. Funke, H. L. Rehm, M. S. Olesen, B. A. Maron, P. Szolovits, D. M. Margulies, J. Loscalzo, and I. S. Kohane. 2016. genetic misdiagnoses and the potential for health disparities. *New England Journal of Medicine* 375:655–665.

Matheny, M., S. Thadaney, M. Ahmed, and D. Whicher, editors. *Artificial Intelligence and Health Care: The Hope, the Hype, the Promise, and the Perils.* Washington, DC: National Academy of Medicine.

Matheson, R. 2018a. Machine-learning system determines the fewest, smallest doses that could still shrink brain tumors. *MIT News*. http://news.mit.edu/2018/artificial-intelligence-model-learns-patient-data-cancer-treatment-less-toxic-0810.

McCarthy, J. 1974. Review of "Artificial Intelligence: A General Survey," by Professor Sir James Lighthill. *Artificial Intelligence* 5:317–322.

Mello, M., and I. Cohen. 2018. HIPAA and protecting health information in the 21st century. *JAMA* 320:231–232.

Meskó, B., G. Hetényi, and Z. Győrffy. 2018. Will artificial intelligence solve the human resource crisis in healthcare? *BMC Health Services Research* 18(1): art. 545.

Miller, R. A., H. E. Pople, and J. D. Myers. 1982. Internist-1, an experimental computer-based diagnostic consultant for general internal medicine. *New England Journal of Medicine* 307:468–476.

Moons, K. G., A. P. Kengne, D. E. Grobbee, P. Royston, Y. Vergouwe, D. G. Altman, and M. Woodward. 2012. Risk prediction models: II. External validation, model updating, and impact assessment. *Heart* 98(9):691–698.

Moor, J. 2006. The Dartmouth College Artificial Intelligence Conference: The next fifty years. *AI Magazine* 27(4):87–91.

Muller, J. Z. 2018. *The Tyranny of Metrics*. Princeton NJ: Princeton University Press.

Murff, H. J., F. FitzHenry, M. E. Matheny, N. Gentry, K. L. Kotter, K. Crimin, R. S. Dittus, A. K. Rosen, P. L. Elkin, S. H. Brown, and T. Speroff. 2011. Automated identification of postoperative complications within an electronic medical record using natural language processing. *JAMA*, 306(8), 848-855.

Newmarker, C. 2018. Digital surgery touts artificial intelligence for the operating room. *Medical Design & Outsourcing*. https://www.medicaldesignandoutsourcing.com/digital-surgery-touts-artificial-intelligence- for-the-operating-room/ (accessed November 12, 2019).

NRC (National Research Council). 1996. *Language and Machines: Computers in Translation and Linguistics: A Report*. Washington, DC: National Academy Press.

Pethokoukis, J. 2016. What the story of ATMs and bank tellers reveals about the "rise of the robots" and jobs. *AEIdeas* Blog. https://www.aei.org/publication/what-atms-bank-tellers-rise-robots-and-jobs/ (accessed November 12, 2019).

Poon, H., C. Quirk, K. Toutanova, and S. Wen-tau Yih. 2018. AI for precision medicine. Project Hanover. https://hanover.azurewebsites.net/#machineReading (accessed November 12, 2019).

Prates, M. O. R., P. H. C. Avelar, and L. Lamb. 2018. Assessing gender bias in machine translation—a case study with Google Translate. *arXiv.org*. https://arxiv.org/abs/1809.02208 (accessed November 12, 2019).

Rycroft-Malone, J. 2004. The PARIHS framework—a framework for guiding the implementation of evidence-based practice. *Journal of Nursing Care Quality* 19(4):297–304.

Schulam, P., and S. Saria. 2017. Reliable decision support using counterfactual models. *Advances in Neural Information Processing Systems* 30 (NIPS 2017), pp. 1697–1708. https://papers.nips.cc/paper/6767-reliable-decision-support-using-counterfactual-models.pdf (accessed November 13, 2019).

Schulam, P., and S. Saria. 2019. Can you trust this prediction? Auditing pointwise reliability after learning. *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1022–1031. http://proceedings.mlr.press/v89/schulam19a/schulam19a.pdf (accessed November 13, 2019).

Sennaar, K. 2018. Machine learning medical diagnostics—4 current applications. Emerj Artificial Intelligence Research. https://emerj.com/ai-sector-overviews/machine-learning-medical-diagnostics-4-current- applications/ (accessed November 12, 2019).

Shah, N. H., A. Milstein, and S. C. Bagley. 2019. Making machine learning models clinically useful. *JAMA* 322(14):1351–1352. https://doi.org/10.1001/jama.2019.10306.

Shanafelt, T. D., S. Boon, L. Tan, L. N. Dyrbye, W. Sotile, D. Satele, C. P. West, J. Sloan, and M. R. Oreskovich. 2012. Burnout and satisfaction with work-life balance among US physicians relative to the general US population. *Archives of Internal Medicine* 172(18):1377–1385.

Shortliffe, E. H. 1974. MYCIN: *A Rule-Based Computer Program for Advising Physicians Regarding Antimicrobial Therapy Selection*. Palo Alto, CA: Stanford Artificial Intelligence Laboratory, Stanford University.

Shortliffe, E. H., and B. G. Buchanan. 1975. A model of inexact reasoning in medicine. *Mathematical Biosciences* 23 (3–4):351–379. doi: 10.1016/0025-5564(75)90047-4.

Sinsky, C., L. Colligan, L. Li, M. Prgomet, S. Reynolds, L. Goeders, J. Westbrook, M. Tutty, and G. Blike. 2016. Allocation of physician time in ambulatory practice: A time and motion study in 4 specialties. *Annals of Internal Medicine* 165:753–760.

Snyder, L., D. McEwen, M. Thrun, and A. Davidson. 2016. Visualizing the local experience: HIV Data to Care Tool. *Online Journal of Public Health Informatics* 8(1):e39.

Soleimani, H., J. Hensman, and S. Saria. 2018. Scalable joint models for reliable uncertainty-aware event prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40(8):1948–1963.

Subbaswamy, A., P. Schulam, and S. Saria. 2019. Preventing failures due to dataset shift: Learning predictive models that transport. *arXiv.org*. https://arxiv.org/abs/1812.04597 (accessed November 14, 2019).

Sumner, W. I. 1993. Review of Iliad and Quick Medical Reference for primary care providers. *Archives of Family Medicine* 2:87–95.

Turing, A. M. 1950. Computing machinery and intelligence. *Mind*, 59:236, 99.49:433–460.

Voosen, P. 2017. The AI detectives. *Science* 357(6346):22–27. https://science.sciencemag.org/content/357/6346/22.summary (accessed November 13, 2019).

Weiner, N. 1948. *Cybernetics: Or Control and Communication in the Animal and the Machine*. Cambridge, MA: MIT Press.

Wiens, J., S. Saria, M. Sendak, M. Ghassemi, V. X. Liu, F. Doshi-Velez, K. Jung, K. Heller, D. Kale, M. Saeed, P. N. Ossorio, S. Thadaney-Israni, and A. Goldenberg. 2019. Do no harm: A roadmap for responsible machine learning for health care. *Nature Medicine*. 25(9):1137-1340.

Wiggers, K. 2018. CB Insights: AI health care startups have raised $4.3 billion since 2013. *VentureBeat*. https://venturebeat.com/2018/09/13/cb-insights-ai-health-care-startups-have-raised-4-3-billion-since-2013/ (accessed November 12, 2019).

Wilkinson, M. D., M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J. W. Boiten, L. B. da Silva Santos, P. E. Bourne, and J. Bouwman. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3:art 160018.

Witten, I. H., E. Frank, M. A. Hall, and C. J. Pal. 2016. *Data Mining: Practical Machine Learning Tools and Techniques*. Burlington, MA, USA: Morgan Kaufmann.

Zauderer, M. G., A. Gucalp, A. S. Epstein, A. D. Seidman, A. Caroline, S. Granovsky, J. Fu, J. Keesing, S. Lewis, H. Co, J. Petri, M. Megerian, T. Eggebraaten, P. Bach, and M. G. Kris. 2014. Piloting IBM Watson Oncology within Memorial Sloan Kettering's regional network. *Journal of Clinical Oncology* 32(15 Suppl):e17653-e17653.

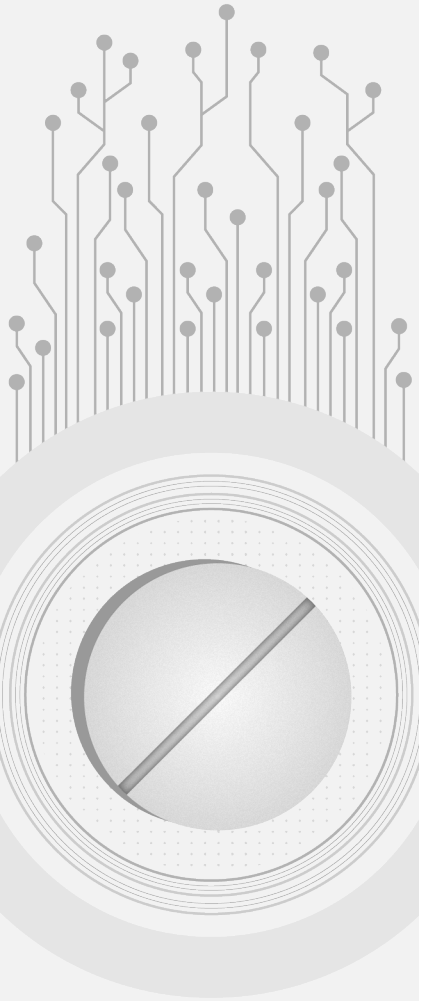
AUTHORS of NAM SPECIAL PUBLICATION

MICHAEL MATHENY, Vanderbilt University Medical Center and the Department of Veterans Affairs (Co-Chair)
SONOO THADANEY-ISRANI, Stanford University (Co-Chair)
ANDREW AUERBACH, University of California San Francisco
ANDY BEAM, Harvard University
PAUL BLEICHER, OptumLabs
WENDY CHAPMAN, University of Melbourne
JONATHAN CHEN, Stanford University
GUILHERME DEL FIOL, University of Utah
HOSSEIN ESTIRI, Harvard Medical School
JAMES FACKLER, Johns Hopkins School of Medicine
STEVE FIHN, University of Washington
ANNA GOLDENBERG, University of Toronto
SETH HAIN, Epic
JAIMEE HEFFNER, Fred Hutchinson Cancer Research Center
EDMUND JACKSON, Hospital Corporation of America
JEFFREY KLANN, Harvard Medical School
RITA KUKAFKA, Columbia University
HONGFANG LIU, Mayo Clinic
DOUG MCNAIR, Cerner
ENEIDA MENDONCA, University of Wisconsin Madison
JONI PIERCE, University of Utah
NICHOLSON PRICE, University of Michigan
JOACHIM ROSKI, Booz Allen Hamilton
SUCHI SARIA, Johns Hopkins University
NIGAM SHAH, Stanford University
RANAK TRIVEDI, Stanford University
JENNA WIENS, University of Michigan

NAM Staff

Development of this publication was facilitated by contributions of the following NAM staff, under the guidance of **J. Michael McGinnis,** Leonard D. Schaeffer Executive Officer and Executive Director of the Leadership Consortium for a Value & Science-Driven Health System:

DANIELLE WHICHER, Senior Program Officer (until September 2019)
MAHNOOR AHMED, Associate Program Officer
JESSICA BROWN, Executive Assistant to the Executive Officer (until September 2019)
FASIKA GEBRU, Senior Program Assistant
JENNA OGILVIE, Deputy Director of Communications



PART TWO

Artificial Intelligence in Health Care:

Benefits and Challenges of Machine Learning in Drug Development

U. S. Government Accountability Office (GAO)

Part Two of this Joint Publication presents the GAO Technology Assessment: *Artificial Intelligence in Health Care: Benefits and Challenges of Machine Learning in Drug Development.* While GAO worked closely with the National Academy of Medicine giving feedback on the January 2019 workshop outputs and in preparing the July 2019 meeting on AI in drug discovery and development, the contents and resulting policy options of this technology assessment are solely those of GAO and are the responsibility of GAO.

Highlights of GAO-20-215SP, a report to congressional requesters

December 2019

TECHNOLOGY ASSESSMENT

ARTIFICIAL INTELLIGENCE IN HEALTH CARE

Benefits and Challenges of Machine Learning in Drug Development

Why GAO Did This Study

Developing and bringing a new drug to market is lengthy and expensive. Drug developers study the benefits and risks of new compounds before seeking Food and Drug Administration (FDA) approval. Only about one out of 10,000 chemical compounds initially tested for drug potential makes it through the research and development pipeline, and is then determined by FDA to be safe and effective and approved for marketing in the United States. Machine learning is enabling new insights in the field.

GAO was asked to conduct a technology assessment on the use of AI technologies in drug development with an emphasis on foresight and policy implications. This report discusses (1) current and emerging AI technologies available for drug development and their potential benefits; (2) challenges to the development and adoption of these technologies; and (3) policy options to address challenges to the use of machine learning in drug development.

GAO assessed AI technologies used in the first three steps of the drug development process—drug discovery, preclinical research, and clinical trials; interviewed a range of stakeholder groups including, government, industry, academia, and nongovernmental organizations; convened a meeting of experts in conjunction with the National Academies; and reviewed key reports and scientific literature. GAO is identifying policy options in this report.

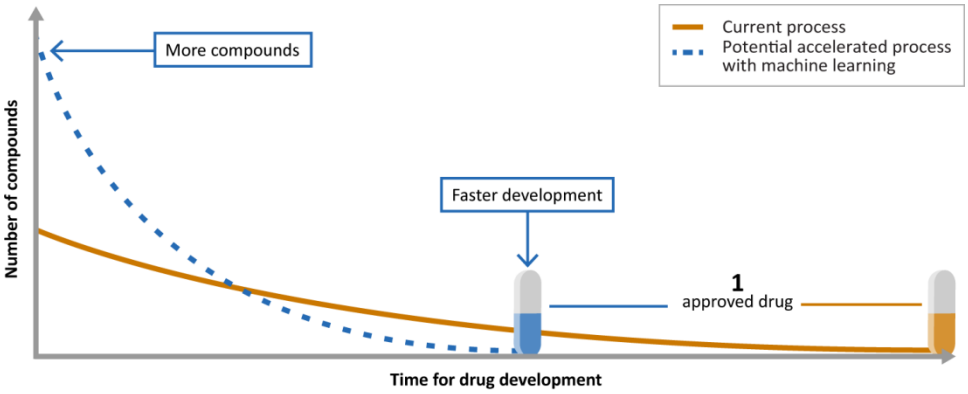
View GAO-20-215SP. For more information, contact Timothy M. Persons, PhD, at 202-512-6888 or personst@gao.gov.

What GAO Found

Machine learning—a field of artificial intelligence (AI) in which software learns from data to perform a task—is already used in drug development and holds the potential to transform the field, according to stakeholders such as agency officials, industry representatives, and academic researchers. Machine learning is used throughout the drug development process and could increase its efficiency and effectiveness, decreasing the time and cost required to bring new drugs to market. These improvements could save lives and reduce suffering by getting drugs to patients in need more quickly, and could allow researchers to invest more resources in areas such as rare or orphan diseases.

Machine learning could accelerate drug development

This set of technologies could screen more chemical compounds and zero in on promising drug candidates in less time than the current process.



Source: GAO. | GAO-20-215SP

Examples of machine learning in the early steps of drug development include:

- **Drug Discovery**: Researchers are identifying new drug targets, screening known compounds for new therapeutic applications, and designing new drug candidates, among other applications.
- **Preclinical Research**: Researchers are augmenting preclinical testing and predicting toxicity before testing potential drugs in humans.
- **Clinical Trials:** Researchers are beginning to improve clinical trial design, a point where many drug candidates fail. Their efforts include applying machine learning to patient selection, recruitment, and stratification.

GAO identified several challenges that hinder the adoption and impact of machine learning in drug development. Gaps in research in biology, chemistry, and machine learning limit the understanding of and impact in this area. A shortage of high-quality data, which are required for machine learning to be effective, is another challenge. It is also difficult to access and share these data because of costs, legal issues, and a lack of incentives for sharing. Furthermore, a low supply of skilled and interdisciplinary workers creates hiring and retention challenges for drug companies. Lastly, uncertainty about regulation of machine learning used in drug development may limit investment in this field.

GAO developed six policy options in response to these challenges. Five policy options are centered around research, data access, standardization, human capital, and regulatory certainty. The last is the status quo, whereby policymakers—federal agencies, state and local governments, academic and research institutions, and industry, among others—would not intervene with current efforts. See below for details of the policy options and relevant opportunities and considerations.

Policy Options to Address Challenges to the Use of Machine Learning in Drug Development

	Opportunities	Considerations
Research (report page 60) Policymakers could promote basic research to generate more and better data and improve understanding of machine learning in drug development.	• Could result in increased scientific and technological output by solving previously challenging problems. • Could result in the generation of additional high-quality, machine readable data.	• Basic research is generally considered a long term investment and its potential benefits are uncertain. • Would likely require assessment of available resources and may require reallocation of resources from other priorities.
Data Access (report page 61) Policymakers could create mechanisms or incentives for increased sharing of high-quality data held by public or private actors, while also ensuring protection of patient data.	• Could shorten the length of the drug development process and reduce costs. • Could help companies identify unsuccessful drug candidates sooner, conserving resources.	• Would likely require coordination between various stakeholders and incur setup and maintenance costs. • Improper data sharing or use could have legal consequences. • Cybersecurity risks could increase, and those threats would likely take additional time and resources to mitigate. • Organizations with proprietary data could be reluctant to participate.
Standardization (report page 62) Policymakers could collaborate with relevant stakeholders to establish uniform standards for data and algorithms.	• Could improve interoperability by more easily allowing researchers to combine different data sets. • Could help efforts to ensure algorithms remain explainable and transparent, as well as aid data scientists with benchmarking.	• Could be time- and labor-intensive because standards development typically requires consensus from a multitude of public and private-sector stakeholders. This process can result in standards development taking anywhere from 18 months to a decade to complete and require multiple iterations.
Human Capital (report page 63) Policymakers could create opportunities for more public and private sector workers to develop appropriate skills.	• Could provide a larger pool of skilled workers for agencies, companies, and other research organizations, allowing them to better leverage advances in the use of machine learning in drug development. • Interdisciplinary teamwork could improve as workers with different backgrounds learn to better communicate with one another.	• Data science-trained workers could exit the drug development field in search of higher-paying opportunities. • Would likely require an investment of time and resources. Companies and agencies will need to decide if the opportunities and challenges justify the investment or shifting of existing resources and how best to provide such training.
Regulatory Certainty (report page 64) Policymakers could collaborate with relevant stakeholders to develop a clear and consistent message regarding regulation of machine learning in drug development.	• Could help increase the level of public discourse surrounding the technology and allow regulators and the public to better understand its use. • Drug companies could better leverage the technology if they have increased certainty surrounding how, if at all, regulators will review or approve the machine learning algorithms used in drug development.	• Would likely require coordination within and among agencies and other stakeholders, which can be challenging and require additional time and costs. • If new regulations are promulgated, compliance costs and review times could be increased.
Status Quo (report page 65) Policymakers could maintain the status quo (i.e., allow current efforts to proceed without intervention).	• Challenges may be resolved through current efforts. • Companies are already using machine learning and may not need action from policymakers to continue expanding its use.	• The challenges described in this report may remain unresolved or be exacerbated.

Source: GAO.

Abbreviations

AI	artificial intelligence
FDA	Food and Drug Administration
HIPAA	Health Insurance Portability and Accountability Act of 1996
IND	investigational new drug application
MELLODDY	Machine Learning Ledger Orchestration for Drug Discovery
MLPDS	Machine Learning for Pharmaceutical Discovery and Synthesis Consortium
NDA	new drug application
NCATS	National Center for Advancing Translational Sciences
NIH	National Institutes of Health
R&D	research and development

U.S. GOVERNMENT ACCOUNTABILITY OFFICE

441 G St. N.W.
Washington, DC 20548

Introduction

December 20, 2019

Congressional Requesters

It can take 10 to 15 years and high costs to develop a new drug and bring it to market.[5] During this time, drug developers conduct tests to study the benefits and risks of new compounds before seeking Food and Drug Administration (FDA) approval. About one out of every 10,000 chemical compounds initially tested for their drug potential makes it all the way through the research and development (R&D) pipeline, and is then determined by FDA to be safe and effective and approved for marketing in the United States. Although high costs and failure rates make drug development risky, creating a safe and effective new drug can be extremely rewarding for both the developer and the public. A highly successful new drug can cure or alleviate diseases affecting millions of people, as well as generate significant revenue—some of which could support R&D on new treatments for other diseases.

We reported in 2006 the view of some stakeholders that drug industry innovation had stagnated.[6] Ten to twenty years ago it was widely recognized that the number of new drugs being produced was generally declining, while R&D expenses were steadily increasing. For example, we found that the drug industry had reported substantial increases in annual R&D costs, and that the number of new drug applications (NDA) approved by FDA had not been commensurate with those investments. At that time, a variety of factors were contributing to the declining productivity of pharmaceutical R&D, according to experts, including limitations on the scientific understanding needed to translate chemical and biological discoveries into safe and effective drugs, business decisions, regulatory uncertainty, and intellectual property issues.

Recent technological developments are bringing new hope to drug development. Advances such as the sequencing of the human genome and the increasing adoption of electronic health records have generated vast amounts of data that could assist in the search for drugs to prevent, treat, or cure serious illnesses. Advanced analytical capabilities, such as machine learning and related artificial intelligence (AI) technologies, are enabling the industry to convert these large volumes of complex data into new insights.

---

[5]For example, one study estimated average out-of-pocket cost per new compound that received FDA approval between 2005 and 2013 to be $1.4 billion. See J. A. DiMasi, H. G. Grabowski, and R. W. Hansen, "Innovation in the Pharmaceutical Industry: New Estimates of R&D Costs," *Journal of Health Economics*, vol. 47 (2016). Other studies suggest lower development costs. For example, another study estimated a median cost to develop cancer drugs of $600 million. See V. Prasad and S. Mailankody, "Research and Development Spending to Bring Single Cancer Drug to Market and Revenues After Approval," *JAMA Internal Medicine*, published online September 11, 2017, https://jamanetwork.com/journals/jamainternalmedicine/article-abstract/2653012.

[6]GAO, *New Drug Development: Science, Business, Regulatory, and Intellectual Property Issues Cited as Hampering Drug Development Efforts*, GAO-07-49 (Washington, D.C.: Nov. 17, 2006).

In view of the potential of AI to help address challenges in drug development, you asked us to conduct a technology assessment in this area with an emphasis on foresight and policy implications. This report discusses (1) current and emerging AI technologies available for drug development, including discovery through clinical trials, and their potential benefits; (2) challenges to the development and adoption of these technologies; and (3) policy options to address challenges to the use of machine learning in drug development.

To address these objectives, we assessed available and developing AI technologies that companies could use during the drug development process as well as the benefits and challenges associated with their use. To do so, we reviewed key reports and scientific literature describing current and developing technologies; attended relevant technical conferences and workshops; and interviewed a variety of stakeholders, including agency officials, drug companies—both biopharmaceutical and machine learning-focused, academic researchers, and nongovernmental organizations.

In addition, we collaborated with the National Academies to convene a 2-day meeting of 19 experts on current and emerging machine learning technologies for use in drug development. We worked with National Academies staff to identify experts from a range of stakeholder groups including federal agencies, academia, industry, and legal scholars, with expertise covering all significant areas of our review. During this meeting, we moderated discussion sessions on several topics related to machine learning in drug development, including research and example technologies; economic, legal, social, and health factors; and policy and regulatory implications. Following the meeting, we continued to seek the experts' advice to clarify and expand on what we heard. Consistent with our quality assurance framework, we provided the experts with a draft of our report and solicited their feedback, which we incorporated as appropriate.

We limited the policy options included in this report to those that met the policy objective and were within the report scope. We present six policy options in response to the challenges identified during our work and discuss potential opportunities and considerations of each. The options are not intended to be inclusive of all potential policy options. To develop the policy options, we prepared a list of potential policy ideas based on a literature search, stakeholder interviews, and the expert meeting. We removed ideas that were not likely to achieve the policy objective or did not fit into the overall scope of our work. We grouped the remaining ideas based on themes (e.g., human capital, data access). We combined ideas that (1) were duplicative, (2) could be subsumed into a higher-level policy option, or (3) were examples of how to implement a policy option rather than the option itself.

We focused our review on selected technologies in the first three steps of the typical drug development process: drug discovery, preclinical research, and clinical trials. We did not assess all available or developing technologies; instead, we selected examples to demonstrate the breadth of machine learning technologies in drug development.

We conducted our work from February 2019 through December 2019 in accordance with all sections of GAO's Quality Assurance Framework that are relevant to technology assessments. The framework requires that we plan and perform the engagement to obtain sufficient and appropriate evidence to meet our stated objectives and to discuss any limitations to our work. We believe that the information and data obtained, and the analysis conducted, provide a reasonable basis for any findings and conclusions in this product.

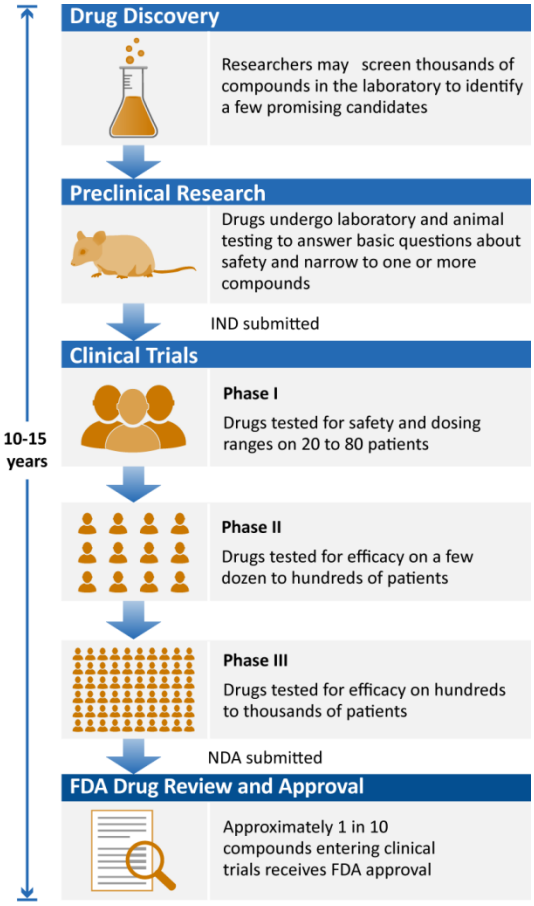
1 Background

1.1 The drug discovery, development, and approval process

FDA is responsible for ensuring the safety and efficacy of drugs marketed in the United States.[7] According to FDA officials, if AI technology is submitted to the agency to support a drug development program or marketing application, FDA will consider that technology in its review process. Additionally, FDA is closely tracking the use of AI in drug development and is considering its policy approach to this area, according to officials.

The process of bringing a new drug to market is long and costly, with only a small fraction of compounds identified early in the process eventually receiving FDA approval (see fig. 1).

Figure 1: The typical drug development and approval process



Source: GAO analysis of Food and Drug Administration (FDA) and Pharmaceutical Research and Manufacturers of America (PhRMA) documentation. | GAO-20-215SP

Note: IND=investigational new drug application, NDA=new drug application. According to FDA officials, there can be wide variation in the number of patients involved in the different clinical trial phases. When a new drug is being tested for a life-threatening ailment, they said, the drug development process may be expedited by going through only one or two phases of clinical trials before an application is submitted to FDA for marketing approval.

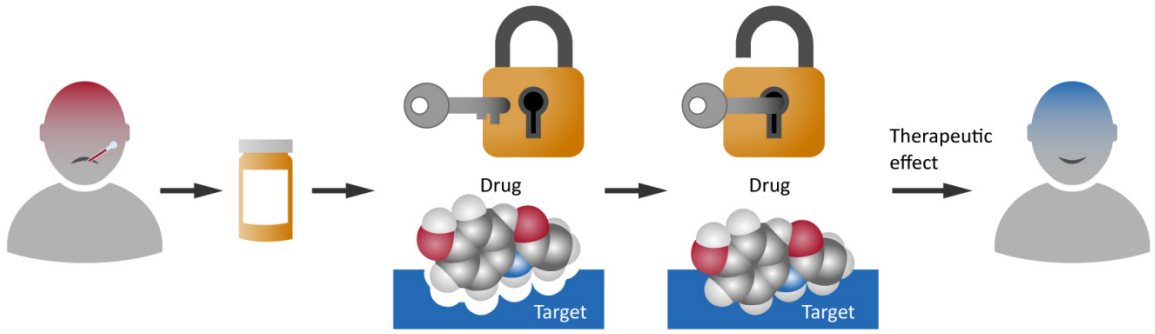
[7]21 U.S.C. § 393(b)(2)(B). Drugs are defined to include, among other things, articles intended for use in the diagnosis, cure, mitigation, treatment, or prevention of disease in man or other animals, and include components of those articles. See 21 U.S.C. §§ 321(g)(1)(B), (D). FDA is also responsible for ensuring the safety, purity, and potency of biological products. 42 U.S.C. § 262(a). Biological products (referred to as biologics in this report) are materials, such as viruses, therapeutic sera, toxins, antitoxins, vaccines, or analogous products to prevent, treat, or cure human diseases or injuries. See 42 U.S.C. § 262(i); 21 C.F.R. § 600.3(h). Most biologics are complex mixtures, and are derived from living sources (such as humans, animals, and microorganisms), unlike most drugs, which are chemically synthesized. Though the FDA approval process is different for drugs and biologics, for the purposes of this report, we refer to drugs and biologics collectively as "drugs."

The drug development process typically consists of the following five steps:[8]

Drug discovery: The drug development process begins with basic research aimed at acquiring new knowledge without immediate commercial application or use. Researchers conduct basic research to better understand the underlying mechanisms of disease, thus increasing the potential for discovering and developing drugs. Using this knowledge, researchers seek to identify and validate a biological target associated with the disease of interest. A target may be a protein, gene, or other biological entity that can be acted on by a drug to achieve a desired therapeutic outcome. The relationship between target and small molecule drug is often described using a lock and key analogy—the drug must fit appropriately into the binding pocket of the target to have an effect (fig. 2).

Good target identification and validation enables increased confidence in the relationship between target and disease. Once a target is identified and validated, researchers screen thousands of compounds from known chemical libraries, or design new compounds, for the desired biological response to the target during testing. Researchers only focus on a small number of these compounds, which have shown the most effective response against the target, to further develop as a potential drug. They conduct experiments to gather information on mechanisms of action, side effects, dosage, delivery, and differential effects across populations, among others, before advancing promising compounds to preclinical research.

Figure 2: The lock and key analogy for drug-target interactions



Source: GAO analysis of the scientific literature. | GAO-20-215SP

[8]This report focuses on the first three of these steps.

Preclinical research: Before testing a drug candidate in humans, drug companies test for toxicity—whether the drug candidate is likely to be safe in humans—using *in vitro* (i.e., in cells or tissues in test tubes or other chambers) and *in vivo* (i.e., in animals) methods. These tests are also used to gather basic information on the safety and efficacy of the drug. If the results are promising, the company may decide to move the drug candidate forward to the next step—clinical trials in humans. Generally, before doing so, the company must submit an investigational new drug application (IND) to FDA; an IND must include, among other things, information from preclinical research and the clinical trial protocols.[9]

Clinical trials: Clinical trials test drug candidates in human volunteers to gather data on safety and efficacy in humans. Typically, clinical trials proceed through phases I, II, and III, generally beginning with testing in a small group of healthy volunteers and then moving on to testing in larger groups of patients the drug candidate is intended to treat. Each clinical trial phase is designed to accomplish something different.[10]

FDA drug review and approval: In most cases, to market a new drug in the United States, drug companies submit an NDA to FDA, which includes safety and efficacy data collected during clinical trials. FDA then reviews and approves the drug for marketing if the data show it to be safe and effective for its intended use.

Post-approval: After FDA has approved a drug and the company has begun marketing, FDA continuously monitors the safety of the drug. FDA can require companies to conduct post-approval studies or clinical trials (known as phase IV clinical trials) to assess a known serious risk, signals of a serious risk, or to identify an unexpected serious risk when data indicate a risk potential.[11] Drug companies may also undertake these studies independently to identify modifications to the drug, such as new delivery mechanisms or additional indications for use.

1.2 Machine learning in AI innovation

Machine learning systems are a central focus of the current AI innovation in drug development.[12] As explained in our previous work, AI has been conceptualized as having three waves of development (see fig. 3):[13]

- Wave 1 – expert or rules-based systems;
- Wave 2 – statistical learning and perceiving and prediction systems; and
- Wave 3 – abstracting and reasoning capability, including explainability.

[9]FDA reviews the IND to, among other things, assure the safety and rights of volunteers who participate in clinical studies. In general, clinical studies may begin 30 days after the FDA receives the IND, unless FDA objects. 21 U.S.C. §§ 355(i)(2), (i)(3); 21 C.F.R. § 312.40.

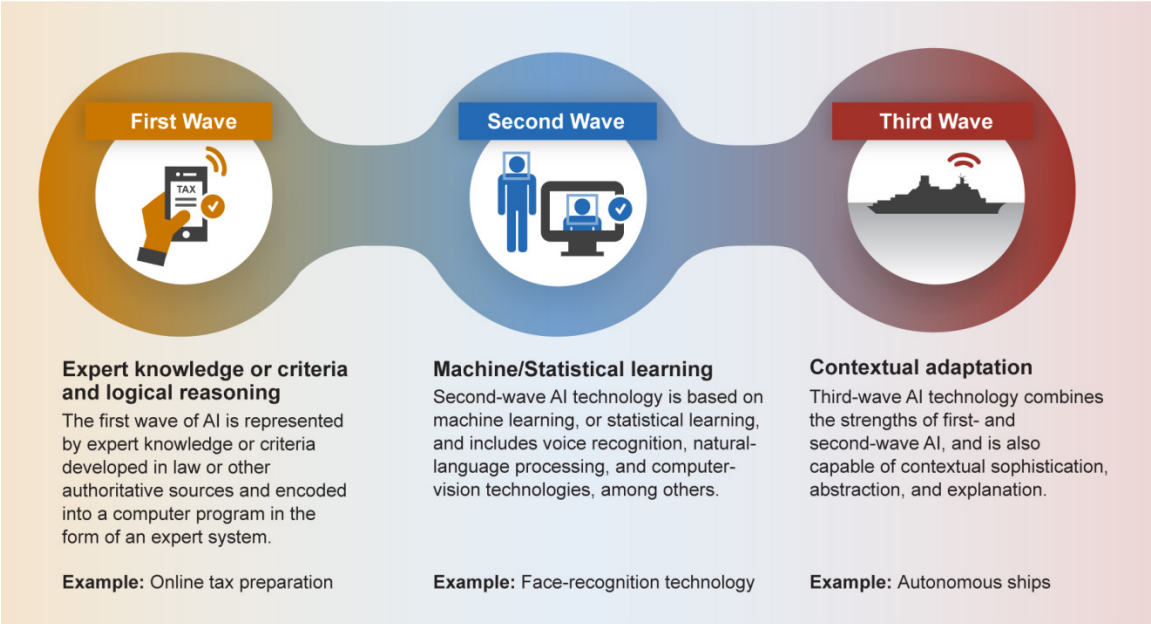
[10]See 21 C.F.R. § 312.21.

[11]21 U.S.C. § 355(o)(3).

[12]AI, which was founded on the idea that machines could be used to simulate human intelligence, has been defined in a variety of ways. Researchers have also distinguished between narrow AI—applications that provide domain-specific expertise or task completion, and general AI—systems that exhibit intelligence comparable to that of a human, or beyond.

[13]GAO, *Artificial Intelligence: Emerging Opportunities, Challenges, and Implications,* GAO-18-142SP (Washington, D.C.: March 28, 2018).

Figure 3: The Three Waves of AI



Source: Defense Advanced Research Projects Agency (DARPA) information; Art Explosion (art). | GAO-20-215SP

Machine learning, the basis for second-wave AI technology, begins with data—generally in vast amounts—and infers rules or decision procedures that accurately predict specified outcomes on the basis of the data provided. In other words, machine learning systems can learn from data, known as the training set, in order to perform a task. Increased availability of large data sets and computing power has enabled recent machine learning advances such as voice recognition by personal assistants on smart phones, an example of natural language processing, and image recognition, an example of computer vision.

Researchers use several methods to train machine learning algorithms, including:

- Supervised machine learning—the data scientist presents an algorithm with labeled data or input; the algorithm identifies logical patterns in the data and uses those patterns to predict a specified answer to a problem. For example, an algorithm trained on many labeled images of cats and dogs could then classify new, unlabeled images as containing either a cat or a dog.
- Unsupervised machine learning—the data scientist presents an algorithm with unlabeled data and allows the algorithm to identify structure in the inputs, for example by clustering similar data, without a preconceived idea of what to expect. In this technique, for example, an algorithm could cluster images into groups based on similar features, such as a group of cat images and a group of dog images, without being told that the images in the training set are those of cats or dogs.
- Semisupervised learning—the data scientist provides an algorithm with a

training set that is partially labeled. The algorithm uses the labeled data to determine the pattern and apply labels to the remaining data.

- Reinforcement learning—an algorithm performs actions and receives rewards or penalties in return. The algorithm learns by developing a strategy to maximize rewards.

While classical machine learning algorithms have been used in drug development for years, recent interest in this area stems from advances in deep learning.[14] An artificial neural network is a machine learning algorithm which, inspired by the brain, contains an input layer that receives data, hidden layers that process data, and an output layer. Deep learning uses deep neural networks, which contain a large number of hidden layers. By contrast, classical artificial neural networks were technologically limited to one or two hidden layers. The types of deep neural networks that are seeing success in other applications are also finding uses in drug development. For example:

- Techniques that are widely used in computer vision can also be used to process biological images such as images of cells from microscopes.
- Techniques often used with sequential data—for example, natural language processing of a text document—can be used to mine scientific literature or process molecular data such as the chemical code in a molecule of DNA.
- Unsupervised learning techniques can be used to generate new chemical structures with desirable therapeutic properties.

Deep learning algorithms, as well as many classical machine learning algorithms, are considered black-box systems, meaning users are unable to understand why the system makes a specific decision or recommendation, why a decision may be in error, or how an error can be corrected. Researchers are actively investigating ways to increase the interpretability or explainability of these algorithms.

1.3 Data generated and used in health care

The generation, collection, access to, and use of data are important aspects of both health care and machine learning research and applications.[15] There are multiple types of data relevant to drug development, including data generated through biomedical research to better understand the biology of diseases and pharmacology of potential drugs, and the various forms of patient data generated in the health care field. Biomedical research data, such as data on the toxicity of known compounds or structures of proteins, may be owned by the organization that generated it or may be publicly available. Recent patient data can be found, for example, in electronic health records, which are digital versions of medical records that can include a person's medical and treatment history, such as diagnoses, medications, and treatment

[14]In this report, we use the term "classical machine learning" to refer to methods not based on deep learning. Examples of classical machine learning include support vector machines and random forest.

[15]This report discusses, generally, the many types of data that could potentially be used for machine learning in drug development. Specific identification of the legal framework that governs each type of data is outside the scope of this report.

plans.[16] Both biomedical research and patient data can be useful for machine learning training, algorithm design, and drug development. However, the factors affecting use of these data differ.

Privacy protections concerning use of health data

Health data, such as the types of patient data described above, may include individually identifiable health information[17] that may be protected by the Health Insurance Portability and Accountability Act of 1996 (HIPAA) and its implementing regulations, known as the Privacy Rule, as well as by other federal and state laws.[18] The HIPAA Privacy Rule governs the use and disclosure of individuals' health information and also provides individuals with privacy rights with regard to their health information. The Privacy Rule generally prohibits regulated entities,[19] which may include health care data warehouses, from using or disclosing protected health information except as specifically permitted, such as for research purposes, under the following conditions:

- with individual authorization,[20]
- without individual authorization if the covered entity obtains documentation that an institutional review or a privacy board has granted waiver of the authorization requirement,[21]
- for review preparatory to research,[22]
- for a limited data set with a data use agreement,[23] and
- if the protected health information has been de-identified.[24]

1.4 Economic considerations of drug development

1.4.1 Grants and tax breaks

The federal government supports new drug R&D both directly—through grants from agencies such as the National Institutes of Health (NIH) and National Science Foundation—and indirectly through tax incentives. Specifically, the Internal Revenue Code includes incentives for research-related spending, for example: through two income tax credits—the credit for clinical testing expenses for certain drugs for rare diseases or conditions,[25] and the credit for increasing

[16]Office of the National Coordinator for Health Information Technology (ONC) https://www.healthit.gov/faq/what-electronic-health-record-ehr (accessed 10/17/2019).

[17]"Individually identifiable health information" is health information, including genetic and demographic information collected from an individual, that (1) is created or received by a health care provider, health plan, employer, or health care clearinghouse; (2) relates to the past, present, or future physical or mental health condition of the individual or the provision or payment for health care to the individual, and (3) can be used to identify the individual or with respect to which there is a reasonable basis to believe the information can be used to identify the individual. 45 C.F.R. § 160.103.

[18]The Privacy Rule preempts any contrary state law unless the provision of the state law relates to the privacy of individually identifiable health information and is more stringent than a standard, requirement, or implementation of the Privacy Rule. Accordingly, state laws may also play a role in the area of health data privacy.

[19]One expert noted that health data relevant to drug development may be held in non-HIPAA-covered environments (for example, non-HIPAA research organizations or technology companies that are not acting on behalf of HIPAA-covered entities).

[20]45 C.F.R. § 164.508(a)(1).

[21]45 C.F.R. § 164.512(i)(l)(i).

[22]45 C.F.R. § 164.512(i)(l)(ii).

[23]45 C.F.R. § 164.514(e).

[24]45 C.F.R. § 164.502(d)(2). One expert noted that information that has been de-identified for purposes of HIPAA may nevertheless be re-identifiable, and this is a source of privacy concerns for the large data sets used for machine learning.

[25]See 26 U.S.C. § 45C.

research activities[26]—and through deductions of research and experimental expenditures.[27]

1.4.2 Economic incentives for innovation

We previously found that, revenues, costs, and policy incentives influence drug industry R&D investment decisions, according to studies and industry experts.[28] For example, drug companies may invest more in R&D of drugs with therapeutic effects for a large number of patients rather than those targeting smaller groups because they expect those investments to generate higher future streams of revenue.[29] Higher costs of research and innovation lead companies to seek to reduce costs by, for example, focusing on cheaper clinical trials, modifying existing drugs, and acquiring existing research projects at lower costs.

Policy incentives, such as patent protection and market exclusivities, can also influence investment decisions. Patents and market exclusivity periods are two ways drug companies may recoup their R&D investments by limiting competition for specified periods of time. Typically, early in the R&D process, companies developing a new brand-name drug apply to the U.S. Patent and Trademark Office for a patent on the active ingredient in the drug, among other things.[30] Once the patent is granted, other drug companies are excluded from making, using, or selling the patented aspect of the drug during the patent term.[31] Additionally, market exclusivity is a specified period of time during which FDA generally cannot approve a similar competing version of the drug for marketing. In general, the availability and length of an exclusivity period depends on the type of drug and its approved indication. For example, new chemical entities are eligible for five years of market exclusivity upon FDA approval. However, there is also some evidence in the economics literature to suggest that greater competition may be associated with higher incentives to innovate in certain circumstances.

[26]See 26 U.S.C. § 41.

[27]See 26 U.S.C. § 174.

[28]GAO, *Drug Industry: Profits, Research and Development Spending, and Merger and Acquisition Deals*, GAO-18-40 (Washington, D.C.: Nov. 17, 2017).

[29]However, there are also incentives to develop drugs for small populations, such as the tax credit for clinical testing expenses for certain drugs for rare diseases or conditions mentioned above.

[30]See 35 U.S.C. §§ 111, 154.

[31]Typically, a patent term is 20 years from the date on which the patent application was filed.

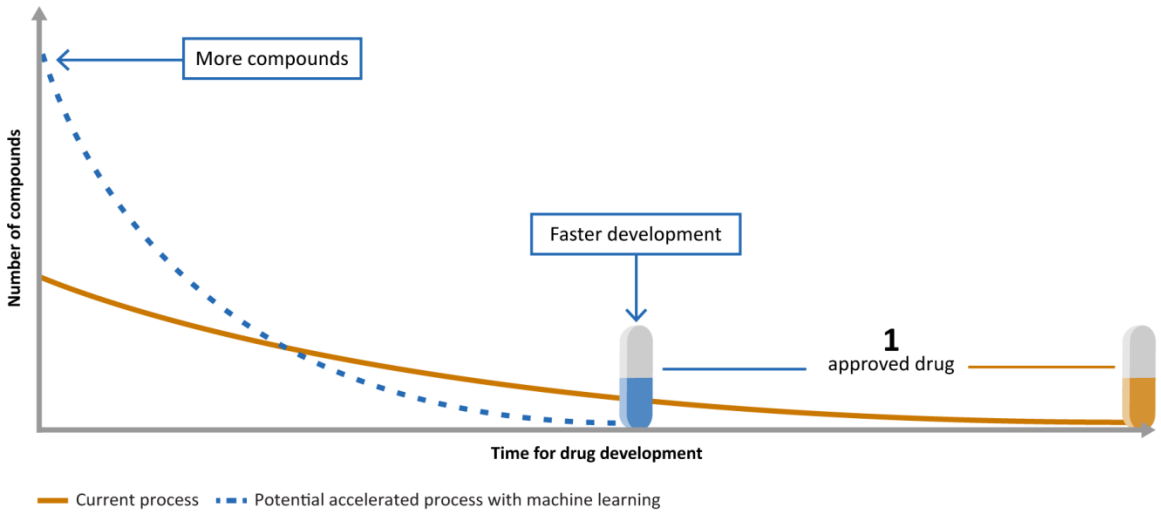
2 Status and Potential Benefits of Machine Learning in Drug Development

Machine learning holds the potential to transform drug development, according to agency officials, industry representatives, and academic researchers. Stakeholders stated that machine learning can make drug development more efficient and effective, decreasing the time and cost required to bring potentially more effective drugs to market (see fig. 4). Both of these improvements could save lives and reduce suffering by getting drugs to patients in need more quickly. Lower R&D costs could also allow researchers to invest more resources in disease areas that are currently not considered profitable to pursue, such as rare or orphan diseases.[32]

Drug companies—including both biopharmaceutical and machine learning-focused companies—are already using machine learning throughout the drug development process. The five biopharmaceutical companies we spoke to said they are incorporating machine learning techniques at every step of the process, while the five machine learning-focused companies we interviewed tend to focus on particular steps or aspects.[33] The best opportunities for machine learning are in particularly data-rich aspects of the process, according to an industry group. For example, natural language processing can help researchers mine the vast and growing scientific literature to prioritize

Figure 4: Machine learning could accelerate drug development

This set of technologies could screen more chemical compounds and zero in on promising drug candidates in less time than the current process.



Source: GAO. | GAO-20-215SP

[32]Rare diseases became known as orphan diseases because drug companies were not interested in adopting them to develop treatments.

[33]One biopharmaceutical company did not directly answer the question but described several examples across the process.

the most relevant publications and identify patterns in published research. According to one biopharmaceutical company representative, these techniques augment the work of their researchers through more focused literature research and by connecting disparate concepts in unanticipated ways.

The specific machine learning techniques that researchers use generally depends on the application. Stakeholders emphasized the importance of using algorithms that are "fit-for-purpose"—that is, using the best algorithm for the specific application. As in other domains, recent advances in deep learning have generated excitement in the field of drug development. These methods offer advantages over classical machine learning, such as the ability to process larger data sets, and can demonstrate improved predictive performance. However, deep learning comes with a high computational cost, and does not always outperform classical machine learning techniques.

Optimizing machine learning algorithms, especially those for deep learning, for specific applications is not trivial, and researchers tend to individualize models for each application or data set. Deep learning also requires large quantities of data, which, according to stakeholders, may not be available depending on the application. For example, according to a company working in clinical trial optimization, they do not use deep learning because the data they use are not available in large enough quantities. According to a scientific review from 2018, it is too early to determine whether deep learning is superior to other machine learning techniques, although deep learning is superior for certain tasks such as image analysis and has shown promise in several applications relevant to drug development.[34]

2.1 Drug discovery

Machine learning in drug discovery, the earliest step of the drug development process, is an active area of research. One company we interviewed said that machine learning during this step will allow them to find better drugs earlier in the process and help speed the drugs through the rest of the process. They estimated that this early-stage acceleration will allow for R&D cost savings of between $300 million and $400 million per successful drug.

Researchers are using machine learning for several purposes in the drug discovery step, including identifying new targets, screening known compounds for new therapeutic applications, and designing new drug candidates.[35] Machine learning techniques are enabling better understanding of disease biology by allowing researchers to mine and analyze large biological data sets. Researchers are, for example, mining biological data to identify new drug targets. Traditionally, researchers often discovered targets through basic research, sometimes relying on a certain level of serendipity. Poor efficacy is a common cause of failure when drug candidates reach clinical trials. One possible reason is a weak biological link between the

[34]H. Chen, O. Engkvist, Y. Wang, M. Olivecrona, and T. Blaschke, "The rise of deep learning in drug discovery," *Drug Discovery Today,* vol. 23, no. 6 (2018).

[35]The examples in this section are not an exhaustive list of machine learning applications in drug discovery. For example, researchers are also employing machine learning for predictive modeling of structure-activity relationships; absorption, distribution, metabolism, and excretion properties; and synthesis planning.

target and the disease the drug candidate is intended to treat.

Machine learning has the potential to find targets with stronger links. For example, researchers recently used a semisupervised learning approach to predict potential drug targets using data on genes that have been found to be associated with various diseases.[36] The data came from Open Targets, a public-private partnership that compiles and openly shares a variety of data that could be used to link genes to diseases for drug target identification and prioritization.[37] The researchers compared four algorithms and found that an artificial neural network performed the best, predicting more than 1,000 genes as potential new targets. Predictive tools such as these could help companies invest resources in targets with a strong link to the disease of interest and therefore increase the likelihood of developing an effective drug. However, such predicted targets still need to be validated in the laboratory, according to the scientific literature. Additionally, while resources such as Open Targets provide open access to biological data, scientific literature indicates that disease-relevant data are generally insufficiently available and not evenly distributed across diseases areas.

One of the most common uses of machine learning in drug development is virtual screening of compounds, according to the scientific literature. There are two principal strategies used in the laboratory for screening compounds: target-based screening, which measures the effect compounds have on a selected target such as a protein, and phenotypic screening, which measures the effect compounds have on a whole system such as a cell or organism. In high throughput screening, automated systems are used to simultaneously run assays, or biological tests, on large numbers of compounds. Pharmaceutical companies often have large, proprietary collections of compounds—known as chemical libraries—for screening. Virtual screening is complementary to high throughput screening and can reduce costs and labor of the screening process. Researchers can use machine learning or other computational techniques to prioritize compounds for further testing, rather than conducting an expensive screen of the full library in the laboratory. Additionally, some of the biological assays discovered through academic research are very good representations of disease phenotypes but are not amenable to high throughput screening, according to an expert. Machine learning can help researchers take advantage of new developments in biology. For example, one expert described how they used three separate machine learning models—trained using data on around 200 compounds found to be active against the target in biological assays—to screen 12 million commercially available compounds for activity on a protein whose dysfunction is associated with heart failure. They purchased 200 compounds based on the results of that screen and found that one-third of those were active in laboratory tests. In contrast, a laboratory screen of a massive chemical library can have a success rate as low as 0.01%. According to the company representative, after further optimization, three compounds advanced to preclinical animal studies. The company completed

[36]E. Ferrero, I. Dunham, and P. Sanseau, "In silico prediction of novel therapeutic targets using gene-disease association data," *Journal of Translational Medicine*, vol. 15, no. 182 (2017).

[37]Data compiled by Open Targets include genetics, gene expression, literature, disease pathway, and drug data.

three design cycles in one year, whereas it could take six years to reach that point in a traditional drug discovery program.

Advances in computer vision due to deep learning have translated into applications in biological imaging, which is widely used throughout the drug development process. According to one expert, it is especially useful in phenotypic drug discovery. Researchers use techniques such as microscopy, among others, to understand the effects of a drug candidate on hosts (humans or animals), organs, tissues, or cells and their organelles. High-throughput microscopy imaging, for example, is often used to screen compounds for biological activity based on the structural changes they induce on a cell or its organelles. According to an expert, the field of biological imaging is rapidly becoming quantitative, presenting an opportunity for data science. Similarly, a company representative told us that microscopy images generate rich information but are hard to interpret, and the use of machine learning could lead to new insights. For example, an expert described how, in a screen for compounds that cause cancer cells to differentiate into a less harmful form, computer vision enabled researchers to precisely measure cell changes in a way that is not possible by the human eye alone. Many studies have demonstrated the ability of deep learning to outperform classical techniques for image analysis in this field, according to a scientific review.[38] Deep learning can process the large amounts of data generated by these and similar techniques and could potentially automate laborious tasks. However, the training of deep neural networks for imaging is time-consuming and computationally expensive, and large, high-quality data sets are relatively rare in biological imaging.

In addition to screening known compounds, researchers may also try to generate new, previously unknown compounds with the desired biological properties, a process known as *de novo* drug design. These efforts have benefitted from advances in deep learning techniques, such as generative adversarial networks and reinforcement learning. These networks contain two models: a generator that creates new molecules and a discriminator that estimates how likely it is that a molecule came from the training set or the generator (see text box). Using reinforcement learning, the generator is rewarded when it fools the discriminator (i.e., the discriminator labels new molecules as coming from the training set), and the discriminator is rewarded when it correctly labels molecules.

Machine learning models can be generative or discriminative

Generator: A generative model learns the distribution of the training set in order to create new, or synthetic, data points. For example, given a training set of molecules with a variety of toxicities, the model generates new molecules with a desired toxicity profile.

Discriminator: A discriminative model learns a direct map from inputs to labels so that it can classify new inputs based on those labels. For example, given the same training set as above, the model will aim to predict the toxicity of a given molecule.

Source: GAO analysis of the scientific literature. | GAO-20-215SP

Models such as these can be used to produce new molecules with desired physical or biological properties. For example, a representative from one machine learning-focused company told us that using generative adversarial networks, they were able to generate very potent inhibitors for a

[38]Chen, Engkvist, Wang, Olivecrona, and Blaschke, "The rise of deep learning in drug discovery," 1248.

particular disease target in less than two months—a process that would normally take two to three years. A potential drawback to this type of technique is that generative adversarial networks and reinforcement learning are prone to the problem of mode collapse, wherein the model only generates a small number of similar solutions. Additionally, researchers must ensure that machine learning-generated molecules are realistic (i.e., chemically stable and able to be synthesized) and should also validate predicted biological or physical properties in the laboratory.

2.2 Preclinical research

Machine learning can be used to augment preclinical testing and predict clinical trial outcomes.[39] While the preclinical step is intended to test for toxicity, according to FDA, researchers also use this step to gather basic information on the safety and efficacy of the drug. Despite these efforts, failure rates in clinical trials remain high. Success during preclinical testing is highly dependent on the selected animal models, which are meant to represent specific aspects of a human disease but cannot reproduce all potential complexities.[40] One expert stated that it may eventually be possible to build machine learning models that are comparable to or better than animal models. According to stakeholders, many animal models are poor predictors of human response to drugs. Computational methods as an alternative or complement to animal models could reduce the time and cost of bringing a new drug to patients by helping better predict clinical trial outcomes and could address concerns about the use of animals in research. However, another expert did not expect the industry to move away from animal studies because of the desire to know whether the drug has any pharmacologically-relevant effects in an actual, complex *in vivo* system before moving to clinical trials.

We recently reported on FDA's efforts to foster the development and evaluation of emerging tools and methods for assessing the safety of FDA-regulated products.[41] FDA issued a roadmap on this topic in December 2017 that does not have an explicit goal to replace, reduce, or refine animal testing but states that new methods may have the potential to do so.[42] In that regard, the roadmap states that FDA will encourage medical product sponsors to submit a scientifically valid approach for using a new method early in the regulatory process and to engage in frequent communication with the agency about the suitability of that method. Previous FDA efforts to promote the use of alternative methods included, for example, 2012 guidance to industry stating that companies may use non-animal alternative methods to test the toxicological safety of

[39]The examples in this section are not an exhaustive list of machine learning applications in preclinical research. For example, researchers are using machine learning to predict clinical efficacy.

[40]T. Denayer, T. Stöhr, M. Van Roy, "Animal models in translational medicine: Validation and prediction," *New Horizons in Translational Medicine*, vol. 2 (2014).

[41]GAO, *Animal Use in Research: Federal Agencies Should Assess and Report on Their Efforts to Develop and Promote Alternatives,* GAO-19-629 (Washington, D.C.: Sept. 24, 2019). FDA-regulated products include human and animal drugs, medical devices, food and food ingredients, and biological and tobacco products.

[42]Food and Drug Administration, *FDA's Predictive Toxicology Roadmap* (Washington, D.C.: December 2017).

pharmaceutical drugs if the methods are appropriate or scientifically justified.[43]

Researchers are also exploring the use of machine learning to predict clinical trial outcomes related to safety. For example, researchers developed a model to predict toxicity before testing drugs in humans.[44] They trained a random forest—a decision-tree-based machine learning model—to distinguish between a list of FDA-approved drugs and a list of drugs that failed for toxicity in clinical trials.[45] The model considers a variety of features, including the drug's molecular properties and drug-likeness, as well as properties of the target, to predict the likelihood of toxicity in clinical trials.[46] According to the researchers, the model is more predictive than some of the other methods currently used to assess toxicity. For example, the model was able to flag several compounds that had been pulled from the market as toxic. However, the majority of clinical trials fail for reasons other than toxicity, such as efficacy or financial reasons. Therefore other factors must be considered to fully predict clinical trial outcomes.

2.3 Clinical trials

Researchers are beginning to use machine learning to improve clinical trial design, a point in the process where many potential drug candidates fail.[47] According to a study by an industry group, on average 9.6 percent of drugs that enter phase I clinical trials ultimately receive FDA approval.[48] For example, if 100 drug candidates entered phase I clinical trials, approximately 63 (63.2%) would advance to phase II clinical trials, with 19 of those (30.7%) advancing to phase III clinical trials. Of those 19 drugs, 11 (58.1%) would advance to the NDA process, and, ultimately, 10 of those (85.3%) would be approved by the FDA (see text box). For clinical trials, companies are still piloting machine learning and are not yet publishing the results, according to an academic research center. The use of AI in clinical trials tends to be less mature than earlier steps in the process because privacy regulations limit the access to and use of patient data, according to an industry group. Clinical trials can be complex and therefore are associated with a significant portion of overall R&D costs, according to a recent review. [49] Several factors in clinical trial design can influence the likelihood of success, including patient selection and recruitment.

[43]Food and Drug Administration, *Guidance for Industry: S6 Addendum to Preclinical Safety Evaluation of Biotechnology-Derived Pharmaceuticals* (Washington, D.C.: May 2012).

[44]K. Gayvert, N. Madhukar, and O. Elemento, "A data-driven approach to predicting success and failures of clinical trials," *Cell Chemical Biology*, vol. 23, no. 10 (2016).

[45]Random forest is an example of a classical machine learning algorithm.

[46]Drug-likeness is a qualitative property of compounds that is a measure of similarity to known drugs.

[47]The examples in this section are not an exhaustive list of machine learning applications in clinical trials. For example, researchers are also employing machine learning for patient adherence and monitoring, and to analyze real world evidence for example data collected via wearable technology.

[48]David W. Thomas et al., *Clinical Development Success Rates 2006-2015* (Washington, D.C.: BIO, 2016).

[49]S. Harrer, P. Shah, B. Antony, and J. Hu "Artificial Intelligence for Clinical Trial Design," *Trends in Pharmacological Sciences*, vol. 40, no. 8 (2019).

There are typically three phases of clinical trials before FDA review

Phase I: This clinical trial phase generally tests the safety of the drug on about 20 to 80 healthy volunteers. The goal of this phase is to determine the drug's most frequent side effects and how it is metabolized and excreted. If the drug does not show unacceptable toxicity in the phase I clinical trials, it may move on to phase II.

Phase II: This clinical trial phase assesses the drug's safety and effectiveness on people who have a certain disease or condition, and typically the assessment is conducted on a few dozen to hundreds of volunteers. Generally, during this phase some volunteers receive the drug and others receive a control, such as a placebo. If there is evidence that the drug is effective and safety data are acceptable in the phase II clinical trials, it may move on to phase III.

Phase III: This clinical trial phase generally involves several hundreds to thousands of volunteers who have a certain disease or condition and gathers more information about the drug's safety and effectiveness, again while being compared to a control.

Source: GAO, *Investigational Drugs: FDA and Drug Manufacturers Have Ongoing Efforts to Facilitate Access for Some Patients*, GAO-19-630 (Washington, D.C.: Sept. 9, 2019) and Food and Drug Administration documentation. | GAO-20-215SP

Selecting and recruiting patients for clinical trials is a complex process, and enrollment challenges are the principal cause for clinical trial delays, according to the scientific literature. One machine learning-focused company described how they use classical machine learning to maximize the probability of a successful clinical trial—meaning FDA approval of the drug—by optimizing a number of design variables, including the number of patients. Usually, clinical trial cohorts are not representative of the general population but rather come from a subpopulation of suitable patients in whom researchers believe they will be able to readily measure drug response. Researchers may consider patients suitable for a number of reasons; for example whether the patient is at the correct stage of disease or has a specific phenotype.[50] Machine learning tools can take advantage of the many kinds of data used to assess suitability, such as genomic data and electronic health record data, which are currently fragmented in different locations and formats. Natural language processing and computer vision are both techniques that could harmonize and analyze these data. However, overfitting of machine learning models is a potential risk if there is an imbalance between different training sets, according to the scientific literature.[51]

Machine learning tools are also helping move clinical research towards precision medicine. Precision medicine, sometimes referred to as stratified medicine, is an emerging approach for disease treatment and prevention that takes into account individual variability in genes, environment, and lifestyle for each person. According to experts, the industry is moving away from blockbuster drugs and instead investigating diseases that are more complex or have smaller patient populations.[52] Patient stratification—a process by which patients are grouped by phenotype or prognosis—is useful in areas that could be considered as clusters of smaller diseases rather than one broad disease, such as oncology and neurology.[53]

[50]A phenotype, in this case, is a set of observable characteristics of a patient produced by his or her genetics interacting with the environment.

[51]Overfitting means that the model performs well on the training set but does not work well on other data sets. It is similar in concept to how humans may overgeneralize about a population based on limited information.

[52]Blockbuster drugs are those that are intended for large patient populations and have the potential to reach $1 billion in annual sales.

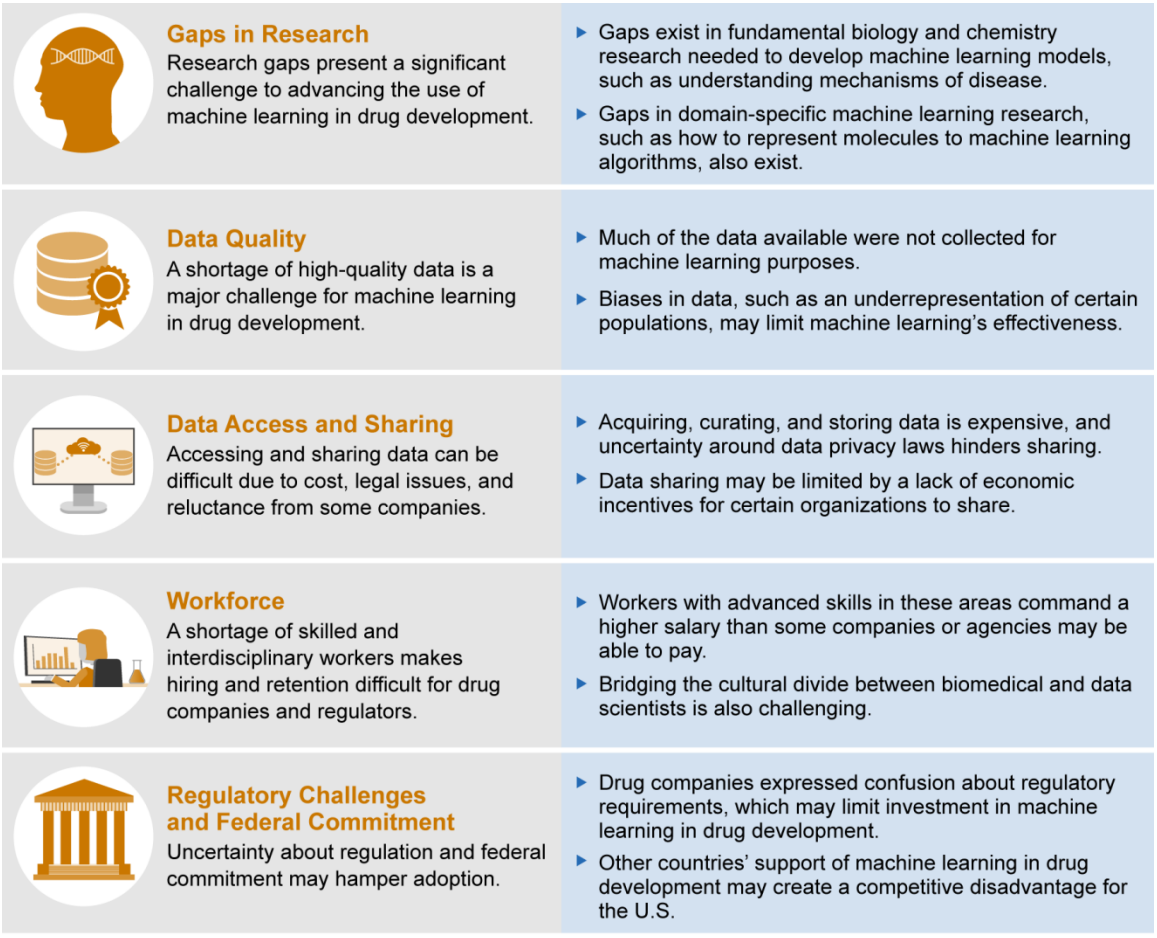
[53]A prognosis is the prospect of recovery for the patient from the disease.

Researchers are using machine learning, both supervised and unsupervised, for patient stratification in clinical trials. Neurology, for example, has one of the highest failure rates among disease areas in clinical trials, according to a study by an industry group. An expert described investigations of the use of machine learning and genetic information to predict whether the rate of cognitive decline in Alzheimer's disease patient subgroups correlates with drug response. In the long term, precision medicine could potentially improve health outcomes through more effective targeted therapies for patient subgroups or individuals. However, certain subpopulations could be excluded from this approach if existing biases in health care data are not overcome. (For more on biases in health care data, see chapter 3.)

3 Challenges Hindering the Use of Machine Learning in Drug Development

Stakeholders, experts, and the literature in this field identified several major challenges hindering the use of machine learning in drug development (see fig. 5). Technological challenges include gaps in the underlying scientific data on mechanisms of disease, structure and behavior of complex molecules, and how to represent these data to algorithms. Stakeholders also point to a shortage of high-quality, unbiased data as well as difficulty accessing and sharing data due to high costs and legal issues. It is also difficult for drug companies to hire and retain skilled, interdisciplinary workers. Finally, regulatory uncertainty and a perceived lack of commitment by the United States compared to other countries can hinder advancement of this field.

Figure 5: Challenges hindering the use of machine learning in drug development



Source: GAO analysis of expert discussions, interviews, and the scientific literature. | GAO-20-215SP

3.1 Gaps in research

Research gaps present a significant challenge to advancing the use of machine learning in drug development. These gaps fall into two broad categories: gaps in understanding of fundamental biology and chemistry, and gaps in domain-specific machine learning research. Experts in the field have noted that addressing these issues may be transformational for future applications of machine learning in drug development.

The federal government has also initiated research into improving predictive screening techniques and other machine learning technologies (see text box).

Existing government research initiatives

Conversations with experts and agency officials uncovered some existing research programs that seek to promote increased development and application of machine learning techniques in biomedical research. According to an official from NIH, broadly, NIH supports research on machine learning, including deep learning, across four primary categories: image analysis, systems pharmacology, predictive screening, and advanced methods development. For example, the NIH National Center for Advancing Translational Sciences (NCATS) developed a funding opportunity that supports the uses of computational algorithms to identify new therapeutic uses of existing drugs and biologics. However, some experts expressed concerns that the current research and training funding model, such as existing NIH study sections and training grants, were not directed appropriately to incentivize the incorporation of machine learning into biomedical research. For example, an expert expressed concern that a lack of machine learning expertise within funding agencies could hinder appropriate review of machine learning-focused research proposals.

Source: GAO. | GAO-20-215SP

3.1.1 Gaps in fundamental biology and chemistry research

- **Understanding mechanisms of disease***:* Experts noted that increased research into the mechanisms of disease could help researchers develop better models which reflect scientific and clinical realities, thus increasing the accuracy of machine learning outputs and target identification in drug development. One research effort that researchers pointed to as a potential model is Genomics England. This project is sequencing the genomes of 100,000 people to identify genetic links to disease. Such research could provide a deeper understanding of disease mechanisms by revealing, for example, what genes and proteins are involved in a disease. This understanding, in turn, could lead to insights into what targets might be suitable for drug development.
- **Modeling drug-protein interactions:** Understanding interactions between proteins and drugs is essential to drug development; however these interactions are complex and not always well understood. For example, drugs interact with multiple systems in the body—a concept called polypharmacology. According to the scientific literature, animal testing and human clinical trials are the current gold standard for testing polypharmacological effects. This may change as computational methods in this field advance. Additionally, target proteins are in rapid, dynamic movement in the body. Computational methods that aim to model static drug-protein interactions therefore have limited accuracy. Experts agreed that there can be instances where diseases occur on the

systems level, and therefore it may not be a single protein that is being regulated but rather multiple proteins or even an entire system or subsystem in the body. More and better data related to these interactions could lead to more accurate deep learning models.

- **Understanding the vast universe of potential compounds:** Experts stated that current chemical libraries contain about 11 billion synthesizable compounds. While this seems like a large number, it is only a tiny fraction of the vast universe of compounds that is theoretically possible. Therefore, any sample of this set of known compounds could be outdated tomorrow and may not provide accurate models of the range of possible compound properties. Furthermore, researchers in the field told us that most data are on small molecules, with very little data on the more complex, large molecules such as biologics. More data on new compounds, both small and large, can reduce the unknown and improve the ability of machine learning algorithms to identify compounds with the best biological response to the target of interest.
- **Understanding why drugs fail:** As noted above, very few drug candidates actually make it to market. More information about why some drug candidates fail to make it to market and others succeed, or whether they have unintended effects, would help researchers develop machine learning algorithms that better predict success of compounds to achieve the desired effect and reduce the time spent on inadequate drug candidates.

3.1.2 Gaps in domain-specific machine learning research

- **The representation of molecules in machine learning:** Researchers can represent molecules to machine learning algorithms in multiple ways, including molecular fingerprints and molecular graphs.[54] Each of these may have advantages and disadvantages, and it is not always clear which representation is the best choice for a given structure. When the researcher selects the type of representation, they are making a subjective choice about the information supplied to the machine learning model, which experts stated will have a significant impact on the resulting output. More research into how the types of molecular representations affect the results of algorithmic selection of compounds could help inform the selection of representations and improve results.
- **Generating new chemical structures:** Additional basic research is needed before machine learning techniques designed to generate new compounds will be fully functional in drug development. For example, as previously discussed, generative models can be used to identify new compounds with the desired biological properties, and predictive models could then be used to evaluate those new compounds as drug candidates. However, those predictive models may not produce reliable results when extrapolated to new compounds

[54]Molecular fingerprints encode structural or functional features of molecules in a binary format and a molecular graph has vertices that represent the atoms and edges that represent the bonds of a particular molecule.

that are significantly different from those on which the models were trained.

- **Unsupervised learning:** Unsupervised learning techniques can yield compelling insights from unlabeled data, such as electronic health records that do not contain patient outcomes. For example, a group of researchers applied an unsupervised deep learning algorithm to 700,000 such electronic health records, and reported that the algorithm was able to consistently and significantly outperform other predictive methods in assessing a patient's future disease profile for the 78 diseases in the study.[55] However, these techniques have not seen widespread use in the biomedical sciences because they can be challenging to use, with a wide range of hit-or-miss results and a need to pre-process the data to remove irrelevant factors. Researchers stated that these obstacles might be overcome with additional research into the predictive elements of certain diseases and better raw representations of the data, such as improved understanding of laboratory results, to improve the algorithm's ability to correlate information and predict future outcomes.

3.2 Data quality

A shortage of high-quality data is a major challenge for machine learning in drug development, according to agency officials and industry representatives. Machine learning requires a large amount of accurate and representative data. This poses a unique challenge in drug development, as much of the data were not originally collected with machine learning in mind and may not be machine-readable or model-ready. Furthermore, according to an industry representative, data collected across different organizations and environments come in different formats, and this lack of standardization in data quality is a barrier. Curating these data is a resource-intensive process, according to stakeholders and the literature.[56] One representative from a drug company told us that 80 percent of their effort goes into accessing and curating data to make it usable for their machine learning applications. The ability to trace data back to the original experiment is also essential for machine learning in drug development, and according to one expert, meaningful machine learning results require an understanding of how the data were collected, used, and analyzed in the research.

Another factor that can reduce data quality is bias, which can skew and limit machine learning outputs. For example, publication bias may result in data skewed toward positive results, as negative results can be less valued and remain unpublished. This may lead to issues when published data are used for machine learning. Similarly, according to one drug company, data from failed clinical trials are often not publicly available and opening this data up could unleash a wave of innovation. In addition, patient data may also be biased as such data are collected mostly from individuals receiving treatment, causing an underrepresentation of data from healthy individuals and an overrepresentation of data

[55]R. Miotto, L. Li, B.A. Kidd, and J.T. Dudley, "Deep Patient: An Unsupervised Representation to Predict the Future of Patients from Electronic Health Records," *Scientific Reports*, vol. 6, no. 26094 (2016).

[56]Curating refers to the process of collecting, organizing, and repurposing data.

from individuals with access to medical care, according to experts. Lastly, data may underrepresent individuals or groups based on race, class, or gender, potentially skewing machine learning models and causing differences in the effectiveness of drugs across subpopulations. Some stakeholders stated that biases may be exacerbated by machine learning, though others said that machine learning can be used to help identify and alleviate these biases.

3.3 Data access and sharing

As mentioned above, drug companies and researchers use data from diverse sources in order to obtain the quantity needed for machine learning to be effective, but accessing and sharing these data can be both costly and challenging. According to one industry representative, collecting data from the early drug discovery phase can be cost prohibitive. This representative said that certain health-related data may cost tens of thousands of dollars, as compared to just cents for other consumer related data that many technology companies use.

Data sharing also presents unique legal issues. According to stakeholders, privacy laws such as HIPAA can make it difficult for drug companies, especially those that are not regulated by HIPAA, to share or access data. One expert, however, said that the privacy laws and regulations may not be the issue but rather their interpretation by organizations that may be hesitant to share data. Two experts also noted that the public is wary of the use of their data for commercial profit. These experts stressed the importance of being transparent with the public about how data will be used and how their data may be used for the greater good. Similarly, experts told us that rules and processes for patient consent to data sharing are complicated and may make it difficult for individuals to give such consent. In addition, one expert cautioned that consent can cause selection biases and could also limit the amount of usable data, but also noted that there are legal pathways for accessing and sharing data for public health purposes, research, and certain other uses.

Lastly, data sharing may be limited by a lack of economic incentives for certain organizations to share. According to a drug company representative and an academic researcher, drug companies consider their data to be valuable, proprietary, and a competitive edge. According to two legal researchers, some drug companies are using mergers and acquisitions to access data because there is no protection or legal framework for the data that are transferred through those transactions.

How partnerships such as the MELLODDY consortium address data sharing challenges

The Machine Learning Ledger Orchestration for Drug Discovery (MELLODDY) project is a consortium and public-private partnership representing 17 partners from pharmaceutical, technology, and academic fields. MELLODDY uses a method called federated learning to train machine learning models across the chemical libraries of 10 drug companies. In federated learning, training data are decentralized. The machine learning model learns from data stored at different geographic locations, ensuring that each drug company's private data set stays within its own secure infrastructure. The consortium uses block chain architecture technology to protect proprietary information while at the same time boosting the predictive performance and applicability of the drug discovery models by leveraging all available data.

Source: GAO analysis of Machine Learning Ledger Orchestration for Drug Discovery documentation. | GAO-20-215SP

3.4 Low supply of skilled and interdisciplinary workers

Experts we spoke with told us that there are not enough highly skilled workers with interdisciplinary expertise across data science and biomedical science. According to one economist, there is a finite supply of workers available to do innovative work in this field. Experts reported challenges with hiring and retention due to competition from larger technology companies that can afford to pay higher salaries. In addition, an expert said that government agencies, including regulators, may have trouble competing for the high-level talent they need to properly understand and regulate drugs developed using these technologies. Experts mentioned the need for alternative and continuous learning education models to keep up with the growing demand for such skills, including reforms to PhD programs, interdisciplinary programs in high school and college, vocational schools, online training, and data science boot camps.

A secondary workforce challenge is the cultural divide between biomedical scientists and data scientists. According to representatives from a drug company, getting groups of different people to engage together and speak in a similar language is challenging. One expert from our meeting said that people from different disciplines often work in siloed teams and stressed the importance of composing interdisciplinary teams with workers from both of these areas.

How one drug company approaches the cultural divide between biomedical and data scientists

A representative from one drug company presented at a conference how they solve the issue of cultural divides between biomedical and data scientists. Incoming data scientists are put through a 3-year job rotation program, where they work in each of the six drug development departments within the company. After the 3-year rotation, the data scientists are then permanently stationed within a department. The representative acknowledged that this approach takes time and means that new data experts in the company are not able to fully immerse themselves in the job for which they were hired until years later. However, they noted that the benefits of increased understanding between the biomedical and data scientists, improved company culture, and increased speed of projects are well worth the lead time.

Source: GAO analysis of conference proceedings. | GAO-20-215SP

3.5 Regulatory challenges and federal commitment

According to several stakeholders, the regulatory process for drugs developed using machine learning is unclear. For example, an industry group and a legal scholar told us that there is not enough guidance about what information or data FDA will require for approval of machine learning uses in the drug development process and that regulatory uncertainty could dissuade drug companies from investing more resources into machine learning in drug development. Another legal scholar we spoke with said that machine learning offers the potential for revolutionary improvements in this field but may require regulatory changes to realize these benefits. For example, if advancements in computational methods prove to be more effective and reliable than animal experimentation, replacing animal testing with machine learning applications may require regulatory changes, according to an

industry representative and two academic researchers. As described earlier in this report, FDA is closely tracking the use of AI in drug development and is considering its policy approach to this area, according to officials.

One representative from a drug company told us that, in their opinion, the United States is not as committed to addressing some of these challenges as other countries, such as South Korea, and China. For example, this representative said China is trying to attract human capital from other countries and is also the owner of the most data for machine learning purposes, which provides a large competitive advantage for Chinese companies. Another example is South Korea's National Cancer Center, which developed the Korea Cancer Big Data Platform, a multi-database framework that collects clinical, genomic, imaging, and biobank data using secure de-identification technology that allows for clinical research and practice in future research.[57]

[57]Hyo Soung Cha, Jip Min Jung, Seob Yoon Shin, Young Mi Jang, Phillip Park, Jae Wook Lee, Seung Hyun Chung, and Kui Son Choi, "The Korea Cancer Big Data Platform (K-CBP) for Cancer Research," *International Journal of Environmental Research and Public Health*, vol. 16, no. 2290 (2019).

4 Policy Options to Address Challenges to the Use of Machine Learning in Drug Development

We identified six policy options in response to the challenges discussed in the previous chapter. Those challenges included gaps in research, data quality concerns, a lack of data access and sharing, a low supply of skilled and interdisciplinary workers, and regulatory challenges. First, we present options that address research, data access, standardization, human capital, and regulatory certainty. Then, we describe how policymakers—Congress, federal agencies, state and local governments, academic and research institutions, and industry, among others—could choose to maintain the status quo. In addition, we discuss potential opportunities and considerations of each option. We focused on policy options that were within the report scope.[58] Policymakers could implement the options in a variety of ways, including by launching a pilot program.

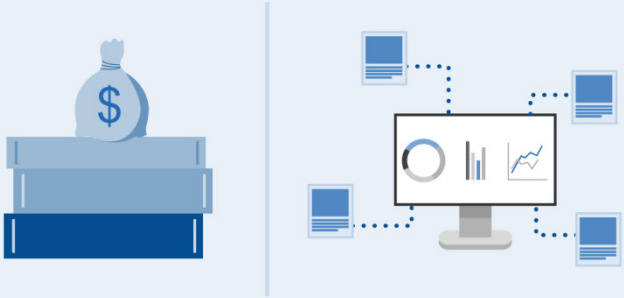
While we present options to address the major challenges we identified, the options are not intended to be inclusive of all potential policy options. We intend policy options to provide policymakers with a broader base of information for decision-making. The options are neither recommendations to federal agencies nor matters for Congressional consideration. They are also not listed in any specific rank or order. We are not suggesting that they be done individually or combined in any particular fashion. Additionally, depending on the options selected, additional work might need to be done on potential design and legal issues. We did not conduct work to assess how effective the options may be, and express no view regarding the extent to which legal changes would be needed to implement them.

[58]For further information on our scope and methodology, please see app. I.

Policy Option: Research

Policymakers could promote basic research to generate more and better data and improve understanding of machine learning in drug development.

Potential Opportunities



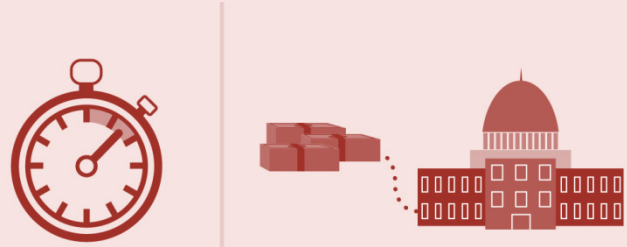
- Could result in increased scientific and technological output by solving previously challenging problems. As we describe above, one such problem is how best to represent molecular structures to machine learning algorithms. Policymakers could promote the field in multiple ways, including approaches such as support for intramural research, grants, or other subsidies. Policymakers could choose to use one of these approaches or combine them.

  In addition, according to one expert, another approach that could promote basic scientific research is a "grand challenge". For example, in 2012, the biopharmaceutical company Merck hosted a challenge for the best predictive model for absorption, metabolism, distribution, excretion, and toxicity modeling of drug candidates.[59]

  Policymakers could also support collaboration across sectors. The Machine Learning for Pharmaceutical Discovery and Synthesis Consortium (MLPDS) is a collaboration between large drug companies such as Pfizer, Merck, and Novartis with the Chemical Engineering, Chemistry, and Computer Science departments at the Massachusetts Institute of Technology, and has published a variety of papers at the intersection of machine learning and drug development.

- Could result in the generation of additional high-quality, machine readable data. For example, as described previously, increased research into the mechanisms of disease could help develop more realistic models.

Potential Considerations



- Basic research is generally considered a long-term investment and its potential benefits are uncertain. For example, the new data created by increased research may not necessarily be high-quality or machine-readable unless data standards are in place.[60]

- Would likely require assessment of available resources and may require reallocation of resources from other priorities.[61] The potential costs borne by any one actor could be mitigated if multiple entities combined their resources.

Source: GAO. | GAO-20-215SP

[59]Merck, *Merck Molecular Activity Challenge*, accessed November 5, 2019, https://www.kaggle.com/c/MerckActivity.

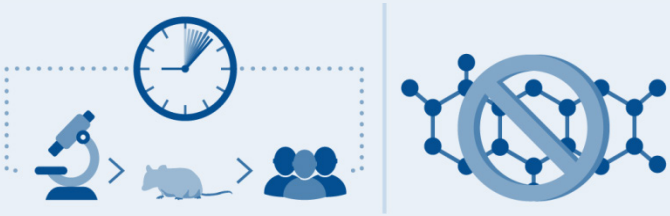
[60]We discuss standardization later in this chapter as its own policy option.

[61]We did not perform an economic analysis to attempt to quantify costs that could be incurred.

Policy Option: Data Access

Policymakers could create mechanisms or incentives for increased sharing of high-quality secured data held by public or private actors, while also ensuring protection of patient data.

Potential Opportunities

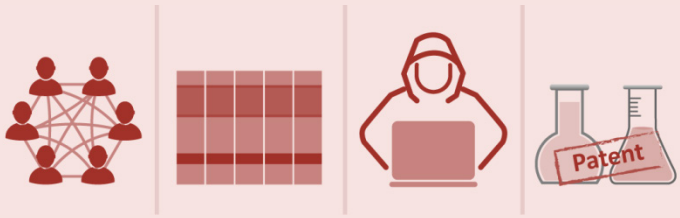


- Could shorten the length of the drug development process and reduce costs. To promote greater availability of data, policymakers could consider forming or facilitating research consortia that allow for secure data sharing. As discussed in chapter 3, the MELLODDY consortium is one example of how different entities are using partnerships to share data. One expert suggested that rewarding researchers based on how often others used their data could incentivize the generation of more open and accessible data. For example, when reviewing grant applications, grant-making organizations could consider how often others have used the applicant researcher's data.

  Policymakers could also consider creating a data repository through encouraging an industry-driven solution, establishing a public-private partnership, or creating a repository of all data under their control. For example, the NIH runs the website ClinicalTrials.gov, which hosts data on over 300,000 research studies. As described earlier, more data could decrease the time to complete clinical trials. However, one of our experts said the website contained data not readily usable for machine learning.

- Could help companies identify unsuccessful drug candidates earlier in the development process, conserving resources. For example, cost reductions could occur within each step of drug development or as a new compound moves from one step to another.[62] As described earlier, studies have estimated the average cost per new FDA-approved drug between $0.6 and $1.4 billion.[63]

Potential Considerations



- Would likely require coordination between various stakeholders and incur setup and maintenance costs.[64] Stakeholders, including federal agencies, would likely need to carefully coordinate across each other's respective domains to minimize duplication and overlap.[65] Previous GAO reports describe how interagency coordination and collaboration can be challenging.[66] For example, a lack of information on roles and responsibilities and lack of coordination mechanisms can hinder effective interagency collaboration. Costs could include computing software and hardware, energy, and staffing needs. Consortia of academic and public entities could combine their efforts to create a repository, spreading the time and cost required across the organizations.
- Improper data sharing or use could have legal consequences. Increased data sharing could therefore require a careful review of the legal ramifications, because data are often gathered through a wide variety of mechanisms and governed by multiple legal frameworks.
- Cybersecurity risks could increase and those threats would likely take additional time and resources to mitigate. For example, if data were stored in a central repository and that system was breached, it could cause a large amount of sensitive data to be exposed at once. In a prior report, we found that cybersecurity breaches in 2015 caused over 113 million individual health care records to be compromised.[67] In May 2015, the University of California, Los Angeles Health network discovered a cyberattack in which individuals had personally identifiable information compromised, including medical record numbers, Medicare or health plan ID numbers, and some medical information.
- Organizations with proprietary data could be reluctant to participate. As we discussed earlier, drug companies consider their data valuable, proprietary, and a competitive edge. Data sharing could raise issues related to ownership or protection of intellectual property. Therefore, incentives may be necessary to encourage the sharing of data. For example, one expert suggested extending patent protection for drugs already developed if drug companies that developed those drugs agreed to release new data.

Source: GAO. | GAO-20-215SP

[62]We did not perform an economic analysis to attempt to quantify cost savings.

[63]DiMasi, Grabowski, and Hansen, "Innovation in the Pharmaceutical Industry," and Prasad and Mailankody, "Spending to Bring Single Cancer Drug to Market and Revenues After Approval."

[64]We did not perform an economic analysis to attempt to quantify costs that could be incurred.

[65]We did not estimate the quantity and quality of coordination needed to implement these policy options.

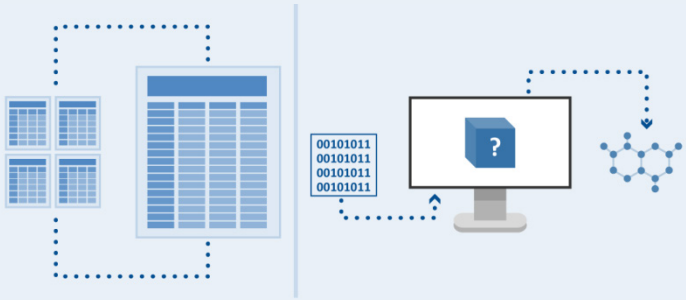
[66]GAO, *Interagency Collaboration: Key Issues for Congressional Oversight of National Security Strategies, Organizations, Workforce, and Information Sharing*, GAO-09-904SP (Washington, D.C.: Sept. 25, 2009). GAO, *Results-Oriented Government: Practices That Can Help Enhance and Sustain Collaboration among Federal Agencies*, GAO-06-15 (Washington, D.C.: Oct. 21, 2005).

[67]This number is based on reported breaches of 500 or more individuals. GAO, *Electronic Health Information: HHS Needs to Strengthen Security and Privacy Guidance and Oversight*, GAO-16-771 (Washington, D.C.: Aug. 26, 2016).

Policy Option: Standardization

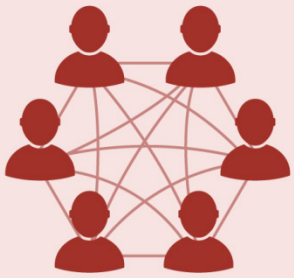
Policymakers could collaborate with relevant stakeholders to establish uniform standards for data and algorithms.

Potential Opportunities



- Could improve interoperability by more easily allowing researchers to combine different data sets. Such combinations can serve the needs of machine learning techniques and reduce bias in the data. Experts we spoke with described the importance of data standardization for their work. For example, a standard that defines synthetic data and how they can be used can help reduce bias by allowing researchers to generate data that could be used to better represent currently underrepresented communities. Similarly, a standard data format for uploading and sharing data across platforms could reduce the need for data scientists to spend time converting data sets to machine-readable formats.
- Could help efforts to ensure algorithms remain explainable and transparent to end users, as well as aid data scientists with benchmarking. Officials from the National Institute of Standards and Technology told us that algorithmic explainability and transparency could encourage adoption of machine learning tools for drug development. The creation of documentary standards could solve such issues by clarifying types of algorithms that can be used for machine learning in drug development.[68] Further, well-established performance standards could help data scientists benchmark their algorithms and make better decisions about when to use certain algorithms over others.

Potential Considerations



- Could be time- and labor-intensive because of the need to reach consensus across a range of public and private-sector stakeholders. Standards development can take anywhere from 18 months to a decade to complete and require multiple iterations.[69] For example, one draft set of private-sector health care AI standards we reviewed began development in March 2019 and is currently on its 14th iteration.[70]

  Standards development organizations follow similar processes and generally adhere to principles such as openness, balance of interests, and consensus.[71] Specifically, once an organization agrees to develop a new or revised standard, it forms a committee of experts from companies, nonprofit organizations, and government agencies. The representatives serve on a voluntary basis, and the committee drafts the standard. Generally, a committee will use a consensus-based process to vote on whether to approve the draft standard. Each step of the process requires careful coordination and collaboration across a myriad of stakeholders.

Source: GAO. | GAO-20-215SP

[68]Standards development organizations can specify how a product is designed or made, or establish performance standards that define the product by function rather than material.

[69]GAO, *National Institute of Standards and Technology: Additional Review and Coordination Could Help Meet Measurement Service Needs and Strengthen Standards Activities*, GAO-18-445 (Washington, D.C.: July 26, 2018).

[70]Consumer Technology Association, *Definitions and Characteristics of Artificial Intelligence in Health Care* (forthcoming).

[71]GAO-18-445.

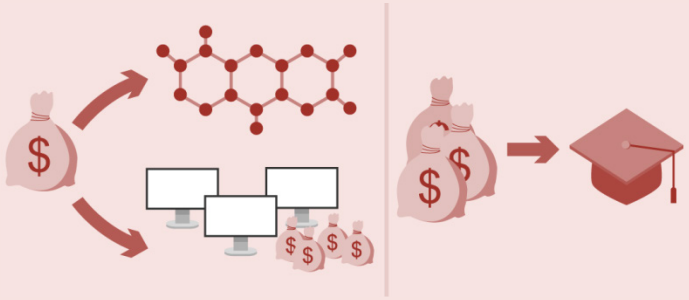
Policy Option: Human Capital

Policymakers could create opportunities for more public and private sector workers to develop appropriate skills.

Potential Opportunities

- Could provide a larger pool of skilled workers for agencies and companies, allowing them to better leverage advances in the use of machine learning in drug development. For example, expanding data science training opportunities for regulators could better equip the agencies to understand machine learning-generated data. Further, if policymakers create opportunities for training workers with relevant knowledge who are motivated to go into the drug development field, companies might be more willing to proceed with efforts to develop machine learning tools for drug development.
- Interdisciplinary teamwork could improve as workers with different backgrounds learn to better communicate with one another. One expert described how interdisciplinary collaboration represented the future of science and noted that discoveries will be made in the space where distinctions between disciplines are blurred. As data scientists learn aspects of biomedical sciences and biomedical scientists learn aspects of data science, the two groups will better navigate different vocabularies and problem-solving approaches. For example, one biopharmaceutical company works to build this collaborative capacity by regularly rotating its workers across multiple divisions of the company to promote interdisciplinary understanding; other organizations could use this type of rotation. Additionally, educational institutions could create interdisciplinary degrees that meld advanced machine learning techniques with biology, chemistry, and medical curricula.

Potential Considerations



- Data science-trained workers could exit the drug development field in search of higher-paying opportunities. We heard from multiple experts and companies that it is extremely difficult to recruit highly qualified interdisciplinary workers because technology companies outside the health care field can offer significantly higher compensation. Companies and agencies might mitigate this concern by offering retention incentives or asking workers to remain with the company for a certain number of years in exchange for the additional training.
- Would likely require an investment of time and resources.[72] Companies and agencies will need to decide if the opportunities and challenges described above justify the investment or shifting of existing resources and how best to provide such training. For example, a company or agency could invest in a partnership with an online teaching platform to create a customized solution rather than developing a new curriculum themselves.

Source: GAO. | GAO-20-215SP

[72]We did not perform an economic analysis to attempt to quantify costs that could be incurred.

Policy Option: Regulatory Certainty

Policymakers could collaborate with relevant stakeholders to develop a clear and consistent message regarding regulation of machine learning in drug development.

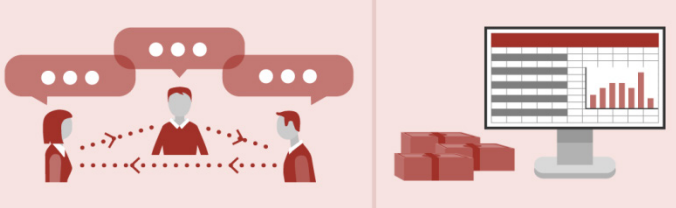
Potential Opportunities

- Could help increase the level of public discourse surrounding the technology and allow regulators and the public to better understand its use. Policymakers can encourage discussion through such mechanisms as holding meetings, issuing discussion papers, engaging with the public, and scientific conferences. For example, in April 2019 FDA released a proposed regulatory framework for modifications to machine learning-based software used as medical devices.[73] The proposal included examples and questions. FDA then gave the public until June 2019 to provide feedback on its proposal. It received over 130 comments from individual citizens, industry groups, health care companies, and one standards association.

  Alternatively, industry could come together to create a self-regulatory mechanism by which it would monitor the use of machine learning in drug development.

- Drug companies could better leverage the technology if they have increased certainty surrounding how, if at all, regulators will review or approve the machine learning algorithms used in drug development. For example, regulators could increase certainty in a variety of ways, including issuance of clarifying documents such as agency guidance or regulations.

Potential Considerations



- Would likely require coordination within and among agencies and other stakeholders, which can be challenging and require additional time and costs.[74] If the process of developing a message is slow, uneven, or inconsistent, industry might lose confidence in the approval process, lessening its interest in pursuing machine learning in drug development. For example, we previously found that inconsistencies among FDA reviewers can influence approval of generic drugs in the first review cycle.[75]

- Compliance costs and review times could increase if new regulations were promulgated, depending on what those regulations require.[76] For example, as new regulations are promulgated, companies might be required to provide additional paperwork, install new or modified capital equipment, or follow new, more rigorous testing procedures. These costs could be absorbed by the companies or passed along to consumers in the form of higher prices. There may also be indirect costs to new regulations, such as a reduction and redirection in industrial R&D efforts.

Source: GAO. | GAO-20-215SP

[73]FDA, "Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD)," Washington, D.C.: Apr. 2, 2019. Medical devices include instruments, apparatuses, machines, and implants that are intended for use to diagnose, cure, treat, or prevent disease, or to affect the structure or any function of the body. See 21 U.S.C. § 321(h).

[74]We did not estimate the quantity and quality of coordination needed to implement these policy options.

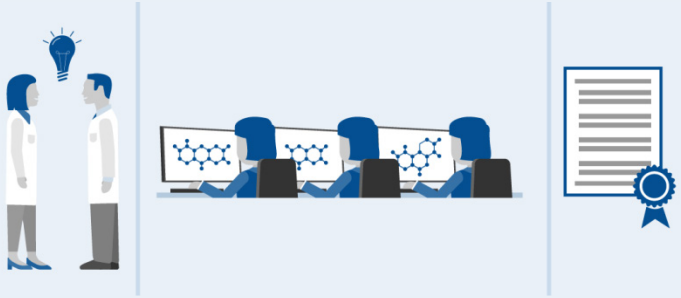
[75]GAO, *Generic Drug Applications: FDA Should Take Additional Steps to Address Factors That May Affect Approval Rates in the First Review Cycle*, GAO-19-565 (Washington, D.C.: Aug. 7 2019).

[76]We did not perform an economic analysis to attempt to quantify costs that could be incurred.

Policy Option: Status Quo

Policymakers could maintain the status quo (i.e., allow current efforts to proceed without intervention).

Potential Opportunities



- Challenges may be resolved through current efforts. For example, with its 2018 Strategic Plan for Data Science, NIH committed to making data findable, accessible, interoperable, and reusable for all data science activities and products supported by the agency. If policymakers allow current efforts time to solve the problems they are targeting, they could avoid potentially spending time and money switching to suboptimal solutions.
- Companies are already using machine learning in drug development and may not need action from policymakers to continue expanding the use of such technologies. As described previously, five biopharmaceutical companies told us they use machine learning throughout the drug development process. Further, five additional companies we spoke with focused on using machine learning in various steps of the drug development process.

Potential Considerations

- The challenges described earlier in the report may remain unresolved or be exacerbated. For instance, the status quo includes the challenge of bias and underrepresentation of certain groups within existing data. Under the status quo, such groups could be further marginalized by the use of such data to develop new drugs that may not be as safe or effective for underrepresented groups.

Source: GAO. | GAO-20-215SP

5 Agency and Expert Comments

We provided a draft of this report to the Department of Commerce (National Institute of Standards and Technology) and Department of Health and Human Services (Food and Drug Administration, National Institutes of Health) with a request for comments. We incorporated agency comments into this report as appropriate.

We also provided a draft of this report to 11 participants from our expert meeting and 2 additional experts for review, and incorporated comments received as appropriate.

As agreed with your offices, unless you publicly announce the contents of this report earlier, we plan no further distribution until 30 days from the report date. At that time, we will send copies of the report to the appropriate congressional committees, relevant federal agencies, and other interested parties. In addition, the report is available at no charge on the GAO website at http://www.gao.gov.

If you or your staff members have any questions about this report, please contact Timothy M. Persons at (202) 512-6888 or personst@gao.gov. Contact points for our Offices of Congressional Relations and Public Affairs may be found on the last page of this report. GAO staff who made key contributions to this report are listed in appendix III.

Timothy M. Persons, PhD

Chief Scientist and Managing Director

Science, Technology Assessment, and Analytics

List of Requesters

The Honorable Lamar Alexander

Chairman
Committee on Health, Education, Labor, and Pensions
United States Senate

The Honorable Greg Walden

Ranking Member
Committee on Energy and Commerce
House of Representatives

The Honorable Michael C. Burgess, MD

Ranking Member
Subcommittee on Health
Committee on Energy and Commerce
House of Representatives

The Honorable Brett Guthrie

Ranking Member
Subcommittee on Oversight and Investigations
Committee on Energy and Commerce
House of Representatives

Appendix I: Objectives, Scope, and Methodology

We describe our scope and methodology for addressing the three objectives outlined below:

1. What current and emerging artificial intelligence (AI) technologies are available for drug development, including discovery through clinical trials, and what are the potential benefits of those technologies?
2. What challenges, if any, hinder the development and adoption of these technologies?
3. What policy options could address challenges to the use of machine learning in drug development?

To address all three research objectives, we assessed available and developing AI technologies that companies could use during the drug development process as well as the benefits and challenges associated with their use. To do so, we reviewed key reports and scientific literature describing current and developing technologies; attended relevant technical conferences and workshops; interviewed a variety of stakeholders, including agency officials, drug companies—both biopharmaceutical and machine learning-focused, academic researchers, and nongovernmental organizations; and conducted an expert meeting in conjunction with the National Academies.

Limitations to scope

We focused our review on selected technologies in the first three steps of the typical drug development process: drug discovery, preclinical research, and clinical trials. Technologies discussed are examples and not an exhaustive list of all AI technologies used in drug development. We did not assess all available or developing technologies. We selected narrative examples to demonstrate the breadth of machine learning technologies in drug development. We also did not include AI technologies for large-scale drug manufacturing.

Literature search

In the course of our work we conducted two literature searches. To establish background and identify appropriate technologies and their benefits and challenges, we reviewed key articles from the scientific literature. To support objective 3, we conducted a policy options literature search using a variety of databases. For this search, results could originate from scholarly or peer reviewed material, government reports, conference papers, dissertations, working papers, books, legislative materials, trade or industry articles, and white papers, but not from general news. We identified a total of 1,109 results using search terms such as “machine learning”, “policy”,“policymaking” and “artificial intelligence”. We selected 116 of the most relevant articles for further review based on our objective, and reviewed the abstracts for additional search terms to refine the results.

Interviews

We interviewed key stakeholders in the field of machine learning in drug development, including representatives or officials from:

- relevant federal agencies including the National Institute of Standards and

Technology, the National Institutes of Health (NIH), and the Food and Drug Administration (FDA);

- five biopharmaceutical companies;
- five companies focused on using machine learning in the drug development process;
- six academic researchers; and
- six nongovernmental organizations including two industry associations, two consumer groups, and two think tanks.[77]

We selected companies to interview by first requesting input from relevant stakeholders. From that initial list, we selected 10 companies that were clearly within the scope of our review and have a U.S. presence. We also balanced our selections between biopharmaceutical and machine-learning focused companies and across the steps of the drug development process. Because this is a small and non-generalizable sample of the universe of companies using machine learning for drug development, the results of our interviews are illustrative and represent important perspectives, but are not generalizable.

Expert meeting

We collaborated with the National Academies to convene a 2-day meeting of 19 experts on current and emerging machine learning technologies for use in drug development. We worked with National Academies staff to identify experts from a range of stakeholder groups including federal agencies, academia, biopharmaceutical companies, machine learning-focused companies, and legal scholars, with expertise covering all significant areas of our review, including individuals with research or operational expertise in using machine learning technology in the drug development process.[78] We evaluated the experts for any conflicts of interest. A conflict of interest was considered to be any current financial or other interest (such as an organizational position) that might conflict with the service of an individual because it could (1) impair objectivity or (2) create an unfair competitive advantage for any person or organization. The 19 experts were determined to be free of reported conflicts of interest, except those that were easily addressed, and the group as a whole was determined to not have any inappropriate biases.[79] (See app. II for a list of these experts and their affiliations.) The comments of these

[77]The six academic researchers focus on the following areas: (1) natural language processing and applications of deep learning to chemistry and oncology, (2) research and development and clinical trial management practices and trends, (3) intellectual property, health law, and regulation, (4) FDA regulation of machine-learning clinical and patient decision support software and gene sequencing and editing technologies; health data privacy and access; genomic civil rights; and citizen science and citizen-led bioethics standard-setting, (5) the intersection of trade secrecy incentives and explainability in AI-enabled health care delivery, and (6) the economics of innovation, intellectual property, productivity measurement, industrial organization, and applied econometrics.

[78]This meeting of experts was planned and convened with the assistance of the National Academy of Science to better ensure that a breadth of expertise was brought to bear in its preparation, however all final decisions regarding meeting substance and expert participation are the responsibility of GAO. Any conclusions and recommendations in GAO reports are solely those of the GAO.

[79]For example, one expert had options in a pharmaceutical company using AI in drug development and sometimes received honoraria for giving talks as part of an advisory board. We determined the expert's relationship did not prevent the expert from serving on the panel as the discussion was not planned to revolve around any specific technology, pharmaceutical company, or vested interest. We did not interview or otherwise mention the company the expert had options in within the report or suggest policy options that will intentionally promote or adversely affect any company.

experts generally represented the views of the experts themselves and not the agency, university, or company with which they were affiliated, and are not generalizable to the views of others in the field.

We divided the 2-day meeting into 7 moderated discussion sessions: (1) technologies to assist with early-stage drug discovery; (2) technologies to assist with drug development; (3) technologies to assist with preclinical research; (4) technologies to assist with clinical trial research; (5) the economic, legal, social, and health factors of AI in drug development; (6) policy and regulatory implications of AI in drug development; and (7) policy options that could facilitate drug development in the United States through the use of AI technologies. Each session featured opening presentations by two to three experts followed by open discussion among all meeting participants based on key questions we provided. The meeting was recorded and transcribed to ensure that we accurately captured the experts' statements. After the meeting, we reviewed the transcripts to characterize their responses and to inform our understanding of all three researchable questions. Following the meeting, we continued to seek the experts' advice to clarify and expand on what we had heard. Consistent with our quality assurance framework, we provided the experts with a draft of our report and solicited their feedback, which we incorporated as appropriate.

Policy Options

We intend policy options to provide policymakers with a broader base of information for decision-making.[80] The options are neither recommendations to federal agencies nor matters for Congressional consideration. They are also not listed in any specific rank or order. We are not suggesting that they be done individually or combined in any particular fashion. Additionally, we did not conduct work to assess how effective the options may be, and express no view regarding the extent to which legal changes would be needed to implement them.

Based on our requesters' interest in U.S. competitiveness and the use of AI technologies in drug development, among other issues, we began our work with an initial policy objective of facilitating drug development in the United States through the use of AI technologies. As our work progressed, we refined this objective to identifying options that could help address challenges to the use of machine learning in drug development. We limited the policy options included in this report to those that met the policy objective and fell within the report scope. We present six policy options in response to the challenges identified during our work and discuss potential opportunities and considerations of each. While we present options to address the major challenges we identified, the options are not intended to be inclusive of all potential policy options.

To develop the policy options, we prepared a list of potential policy ideas (97 in total, as well as the status quo) based on our literature search, stakeholder interviews, and expert meeting. We removed ideas that were not likely to achieve the policy objective or did

[80]Policymakers is a broad term including, for example, Congress, federal agencies, state and local governments, academic and research institutions, and industry.

not fit into the overall scope of our work. For example, we removed policy ideas related to drug pricing that were not relevant to the use of machine learning in the drug development process. We grouped the remaining ideas based on themes (e.g., human capital, data access). We combined those that (1) were duplicative, (2) could be subsumed into a higher-level policy option, or (3) were examples of how to implement a policy option rather than the option itself.

We conducted our work from February 2019 through December 2019 in accordance with all sections of GAO's Quality Assurance Framework that are relevant to technology assessments. The framework requires that we plan and perform the engagement to obtain sufficient and appropriate evidence to meet our stated objectives and to discuss any limitations to our work. We believe that the information and data obtained, and the analysis conducted, provide a reasonable basis for any findings and conclusions in this product.

Appendix II: Expert Participation

We collaborated with the National Academies to convene a two-day meeting of experts to inform our work on artificial intelligence in drug development; the meeting was held on July 18-19, 2019, in Boston, Massachusetts. The experts who participated in this meeting are listed below. Many of these experts gave us additional assistance throughout our work, including seven experts who provided additional assistance during our study by sending material for our review or participating in interviews; and eight experts who reviewed our draft report for accuracy and provided technical comments.

Brandon Allgood

Chief Technology Officer and Cofounder

Numerate, Inc.

Mohammed AlQuraishi

Departmental Fellow in Systems Pharmacology

Harvard Medical School

Regina Barzilay

Professor, Department of Electrical Engineering and Computer Science

Massachusetts Institute of Technology

Anne Carpenter

Institute Scientist and Merkin Institute Fellow

Broad Institute of Harvard and MIT

Will Chen

Head of Computation and Systems Biology

Biogen

Ethan Dmitrovsky

President

Leidos Biomedical Research

Laboratory Director

Frederick National Laboratory for Cancer Research

Shahram Ebadollahi

Global Head of Data Science and AI

Novartis

Olivier Elemento

Director, Englander Institute for Precision Medicine

Weill Cornell Medical School

M. Khair ElZarrad

Deputy Director, Office of Medical Policy at the Center for Drug Evaluation and Research

Food and Drug Administration

Barbara Evans

Director, Center for Biotechnology & Law

University of Houston

Sandy Farmer

Principal

NextGenTech Pharma Consultants

Susan Gregurick

Director, Division of Biomedical Technology, Bioinformatics, and Computational Biology

National Institute of General Medical Sciences, National Institutes of Health

Abraham Heifets

Chief Executive Officer

Atomwise, Inc.

Kenneth I. Kaitin

Professor and Director, Tufts Center for the Study of Drug Development

Tufts University School of Medicine

Michael Keiser

Assistant Professor, Department of Pharmaceutical Chemistry and the Institute for Neurodegenerative Diseases

University of California, San Francisco

Patricia McGovern

Vice President and Global Head of Innovation (Digital) for Regulatory Affairs

Novartis

Arti Rai

Professor and Faculty Director, Center for Innovation Policy

Duke Law

Bobbie-Jo Webb-Robertson

Chief Scientist Computational Biology, Biological Sciences Division

Pacific Northwest National Laboratory

Chris Whelan

Senior Scientist

Biogen

Appendix III: GAO Contact and Staff Acknowledgments

GAO contact

Timothy M. Persons, (202) 512-6888 or personst@gao.gov

Staff acknowledgments

In addition to the contact named above, Karen Howard (Assistant Director), Rebecca Parkhurst (Analyst-in-Charge), Virginia Chanley, Lacey Coppage, Caitlin Dardenne, Leia Dickerson, Bryce Fauble, Matt Hunter, Timothy Kinoshita, Anika McMillon, Jon Menaster, Silda Nikaj, Amanda Postiglione, and Ben Shouse made key contributions to this report. Frederick K. Childers and Katrina Pekar-Carpenter also contributed to this report.

(103352)

GAO's Mission

The Government Accountability Office, the audit, evaluation, and investigative arm of Congress, exists to support Congress in meeting its constitutional responsibilities and to help improve the performance and accountability of the federal government for the American people. GAO examines the use of public funds; evaluates federal programs and policies; and provides analyses, recommendations, and other assistance to help Congress make informed oversight, policy, and funding decisions. GAO's commitment to good government is reflected in its core values of accountability, integrity, and reliability.

Obtaining Copies of GAO Reports and Testimony

The fastest and easiest way to obtain copies of GAO documents at no cost is through GAO's website (https://www.gao.gov). Each weekday afternoon, GAO posts on its website newly released reports, testimony, and correspondence. To have GAO e-mail you a list of newly posted products, go to https://www.gao.gov and select "E-mail Updates."

Order by Phone

The price of each GAO publication reflects GAO's actual cost of production and distribution and depends on the number of pages in the publication and whether the publication is printed in color or black and white. Pricing and ordering information is posted on GAO's website, https://www.gao.gov/ordering.htm.

Place orders by calling (202) 512-6000, toll free (866) 801-7077, or
TDD (202) 512-2537.

Orders may be paid for using American Express, Discover Card, MasterCard, Visa, check, or money order. Call for additional information.

Connect with GAO

Connect with GAO on Facebook, Flickr, Twitter, and YouTube.
Subscribe to our RSS Feeds or E-mail Updates.
Listen to our Podcasts and read The Watchblog.
Visit GAO on the web at https://www.gao.gov/podcast/watchdog.html.

To Report Fraud, Waste, and Abuse in Federal Programs

Contact: Website: https://www.gao.gov/fraudnet/fraudnet.htm
Automated answering system: (800) 424-5454 or (202) 512-7470

Congressional Relations

Orice Williams Brown, Managing Director, williamso@gao.gov, (202) 512-4400, U.S. Government Accountability Office, 441 G Street NW, Room 7125, Washington, DC 20548

Public Affairs

Chuck Young, Managing Director, youngc1@gao.gov, (202) 512-4800
U.S. Government Accountability Office, 441 G Street NW, Room 7149
Washington, DC 20548

Strategic Planning and External Liaison

James-Christian Blockwood, Managing Director, spel@gao.gov, (202) 512-48707
U.S. Government Accountability Office, 441 G Street NW, Room 7814, Washington, DC 20548