The Promise of Data-Driven Drug Development

By Joshua New | September 18, 2019

INTRODUCTION

Overall, policymakers' highest priority should be to dramatically increase the availability of data for drug development.

From screening chemical compounds to optimizing clinical trials to improving post-market surveillance of drugs, the increased use of data and better analytical tools such as artificial intelligence (AI) hold the potential to transform drug development, leading to new treatments, improved patient outcomes, and lower costs. However, achieving the full promise of data-driven drug development will require the U.S. federal government to address a number of obstacles. This should be a priority for policymakers for two main reasons. First, enabling data-driven drug development will accelerate access to more effective and affordable treatments. Second, the competitiveness of the U.S. biopharmaceutical industry is at risk so long as these obstacles exist. As other nations, particularly China, pursue data-driven innovation, especially greater use of AI, foreign life sciences firms could become more competitive at drug development.

Policymakers should recognize that the potential of data-driven drug development is crucial to the well-being of Americans as well as U.S. competitiveness, and develop policies to accelerate this transformation. To that end, policymakers should prioritize data-driven drug development. Overall, policymakers' highest priority should be to dramatically increase the availability of data for drug development—and the most effective way to do that would be to support the creation of a National Health Research Data Exchange to prioritize the collection and sharing of patient medical data for research purposes.

Policymakers should take other steps, including:

- Implementing a unique patient identifier to improve data integrity throughout the health care system;

- Better enforcing the publication of data from clinical trial results by being diligent about penalizing noncompliance;
- Developing guidance for the use of new kinds of data sources in the drug development lifecycle;
- Developing mechanisms to facilitate data sharing by biopharmaceutical stakeholders by establishing data trusts that protect sensitive and proprietary data while still making it available to researchers;
- Requiring and funding the Food and Drug Administration (FDA) to improve the reliability of data used in drug development outside the United States;
- Developing best practices for data collection in health care to ensure equitable outcomes, such as strategies to increase coverage of underrepresented populations;
- Fully funding the U.S. National Institutes of Health (NIH) to accelerate the development of the All of Us Research Program's million-person research cohort; and
- Increasing the number of workers—including high-skilled foreign-born workers—with AI skills and computer science education at all levels.

THE ROLE OF DATA IN THE DRUG DEVELOPMENT LIFECYCLE

There is a wide variety of benefits data can offer for every stage of the drug development lifecycle. This lifecycle can be divided into four distinct stages: discovery, clinical research, FDA review, and FDA post-market safety monitoring.[1] More data, better data, and the increased ability to use data with new technologies such as AI are not only enabling the development of new and more-effective drugs, but are making it easier and more cost-effective do so than ever before.

DISCOVERY

Advances in computing, and large amounts of data, have transformed the field of pharmacology, which in the past primarily relied on the analysis of existing remedies and serendipitous discovery. It has since advanced to primarily using vast reference libraries of molecules and their chemical properties to enable target-based drug discovery, in which researchers engineer drugs based on high-throughput screening (HTS) of molecules known to be involved in a disease process and observing their effects on specific proteins.[2] Thus, as pharmacology is already a data-driven field, the future of pharmacology, which will see access to larger amounts of data

and more powerful analytical techniques—particularly those made possible with AI—should likely entail significant improvement and refinement of existing techniques, rather than a fundamental transformation in how drug discovery is performed.

Academic researchers, pharmaceutical companies, and government agencies have invested huge amounts of time and resources into establishing robust datasets of chemical properties, genetic information, and analytics techniques to enable and improve target-based discovery.[3] These methods have enabled the development of 78 of the 113 first-in-class drugs (drugs that use novel methods of action, or specific biochemical interactions, rather than modified methods of existing therapies) approved by the FDA from 1999 to 2013.[4]

Access to greater computational power and more advanced algorithms has also led to a resurgence in phenotypic screening, which uses the same approach as target-based discovery, but instead observes the effects of a compound in cells, tissues, or even whole organisms, rather than just a protein.[5] Phenotypic screening was the default approach to pharmacology for decades, but did not lend itself well to HTS due to the level of complication involved in the analysis.[6] AI is well suited for this task, and has enabled the large-scale application of HTS to phenotypic screening. For example, a common phenotypic screening method involves using imaging tools to record the fluorescent readout of cell-based assays, in which fluorescent molecules are injected into a sample, and their distribution in a sample can indicate changes in cell function or reaction to a new molecule.[7] However, even automated analysis of these readouts can be a time- and labor-intensive process that still requires manual input and is prone to errors due to the complexity of the analysis involved.[8] Furthermore, performing a new assay requires adjusting and tuning an existing image analysis program, which can take days. A company called Genedata Imagence has developed a deep learning system that can not only substantially reduce the time and effort involved in this fluorescence analysis, but also translate knowledge obtained from one assay to another without human intervention in just minutes.[9] Developing machine learning approaches for this kind of image-based profiling is an active field of research.[10]

Nonetheless, even with the integration of AI, drug discovery is, as one medical researcher noted, still a "lengthy, expensive, difficult and inefficient process with [a] low rate of successful therapeutic discovery."[11] Cost in particular is a major barrier to drug discovery. According to a 2014 study from Tufts University, developing a new drug costs $2.56 billion.[12]

Data-driven technologies have the potential to reduce obstacles to discovery. Automation, involving robotics, sensors, and analytics software, is particularly valuable in HTS due to the large amounts of analysis required to identify potential drug candidates. For example, researchers at Pennsylvania State University used automated screening techniques to analyze over 500,000 chemical compounds to identify a molecule that

could treat or prevent malaria, successfully identifying 631 promising compounds that could potentially be developed into an effective antimalarial drug.[13] However, this screening process still took two years and required the researchers to perform additional analysis to weed out compounds that were toxic before they could identify those 631 candidates.

An even more cost- and time-saving approach to HTS would be to reduce the amount of screening needed to identify drug candidates. Applying AI at various stages of the screening process can enable researchers to make more informed hypotheses about what molecules to screen, enabling HTS to produce more useful results with less screening. For example, a team of researchers led by North Carolina-based Collaborations Pharmaceuticals used machine learning to search through libraries of over 200,000 molecules to identify compounds with desired characteristics for the treatment of tuberculosis.[14] The researchers were able to virtually weed out molecules based on data about their bioactivity and cytotoxicity without ever having to physically screen them, allowing them to screen just 110 molecules in order to identify 2 candidate compounds that can now be the focus of additional research and potentially become new drugs.[15] Such *in silico* approaches to HTS are increasingly common and can make HTS substantially more productive. A typical HTS process has a "hit rate" of identifying candidate molecules of less than 0.05 percent, whereas HTS using only compounds prescreened *in silico* often have hit rates between 10 and 15 percent.[16]

In the future, automation and AI can further reduce the resource demands of drug screening by automating parts of the drug discovery process. In June 2018, the U.K. government announced plans for an automated drug discovery process facility to reduce the time it takes to discover new drugs. The goal of the initiative is to increase the productivity of drug discovery by 5 to 10 times, in part through developing both automated systems that can screen hundreds of thousands of candidate molecules at a time and new, high-resolution imaging technology.[17]

Advances in genetic sequencing, described later in this report, have enabled new techniques for making HTS more cost-effective. Traditional HTS approaches involve discrete analysis of potentially millions of compounds in a reference library to determine whether they affect the function of a target protein. Though analytics-based approaches have made HTS more efficient than in the past, this process can still take large amounts of time and money to complete. In recent years, biopharmaceutical researchers have increasingly relied on DNA-encoded libraries (DELs) to dramatically improve the efficiency of screening. DELs consist of potentially trillions of compounds, each bound to unique DNA strands and stored in several milliliters of solution.[18] Rather than performing thousands or millions of discrete assays, screening with DELs involves simply repeatedly combining this mixture with target proteins and enzymes that wash away any compounds that did not bind with a target. The unique DNA strands on each compound effectively serve as a barcode,

so after an assay is performed, DNA analysis can rapidly identify which compounds successfully bonded with target protein, and then conduct further research on these compounds. DELs have considerable benefits. For example, Danish pharmaceutical company Neuvolution announced in 2017 that it had created a DEL with 40 trillion compounds that is stored in just a couple drops of solution, whereas a traditional HTS reference library of the same size would require vast amounts of space.[19] Additionally, according to pharmaceutical research and development (R&D) company Pharmaron, assembling and screening a library of 1 million compounds costs between $400 million and $2 billion, or approximately $1,100 per compound, whereas a DEL of 800 million compounds costs just $150,000 to assemble and screen.[20] This can make it significantly more cost effective for large firms and start-ups alike to screen compounds that could potentially lead to new drugs.

Advanced computational techniques are also enabling entirely new methods for analyzing the effect of candidate compounds. For example, though proteins are constantly in motion and changing shape, candidate molecule screening often assumes a target protein is in its default state, despite the fact different shapes can alter protein function. A Cambridge, Massachusetts, company called Relay Therapeutics has developed visualization and AI techniques that can analyze this protein motion and screen molecules that can stabilize a protein into its "normal" shape, which could lead to the creation of drugs that are more selective and more effective.[21]

CLINICAL RESEARCH

As with drug screening, a major barrier to developing new treatments is the cost of evaluating candidate drugs for safety and efficacy. As of 2018, the average cost of an individual clinical trial was $19 million.[22] This is consistent with a 2014 study from the Department of Health and Human Services (HHS) that estimated the total costs of Phase I, II, III, and IV trials for a drug to be between $44 million and $115.3 million.[23] Improved use of data and analytics can significantly reduce the costs of clinical trials.

One of the most promising ways to achieve this is the improved use of data and AI in clinical-trial design, particularly to increase patient recruitment and engagement. Selecting a site to perform a clinical trial can be a significant financial commitment, made especially risky due to there being no guarantees patients will actually show up. To minimize this risk, companies such as Trials.ai and Vitrana have developed AI systems that can guide site-selection decisions by analyzing factors such as historical site-performance data and study requirements.[24] A number of companies are using AI to improve patient recruitment directly. For example, Deep 6 AI analyzes structured and unstructured clinical data to better identify patients that match trial criteria, allowing trial organizers to conduct more targeted recruitment.[25] London-based company Antidote uses machine learning for similar purposes, which the company claims enabled it to refer 8,000 patients for a clinical trial relating to Alzheimer's diseases in under

two months—and these referrals were seven times more likely to follow through with the recruitment process than those from other sources.[26]

Even when a trial has enough recruits, to be successful, these recruits must participate in the full trial. However, failure to engage participants properly can cause them to drop out or not adhere to trial rules, thereby reducing the trial's effectiveness. Palo Alto start-up Brite Health has developed a smartphone app that uses machine learning to improve and maintain patent engagement in order to reduce this risk. The app provides users with notifications and nudges them to perform required tasks and site visits. The app also utilizes a chatbot that can make trial information more accessible to patients, while algorithms identify and flag indicators of patient disengagement for trial organizers.[27]

In some cases, patients may end their participation in a trial due to the negative side effects of a treatment. Here, too, AI can help. Researchers have developed machine learning algorithms that can identify the fewest and smallest doses of a chemotherapy regimen that can still shrink brain tumors, thus reducing the toxicity of the treatment.[28] In a simulated trial, the researchers' machine learning model was able to reduce treatment potency by between 25 and 50 percent of all doses without reducing effectiveness.[29] By minimizing the risk of side effects, researchers can more reliably ensure patient adherence to a clinical trial.[30]

New technologies also make it possible to conduct decentralized and virtual clinical trials, which can both make it easier to recruit patients from a wide area and reduce overhead costs. In October 2017, life sciences company AOBiome Therapeutics completed a 12-week clinical trial of an acne drug that proved to be safe and effective.[31] Unlike a traditional clinical trial, however, participants completed the trial at home. AOBiome mailed participants either the drug or a placebo, along with an iPhone that came pre-loaded with an app for participants to take and share regular selfies of their acne, as well as communicate with study organizers throughout the trial.[32] This approach enabled an effective clinical trial with no in-person screening or site visits, which substantially reduced both costs and barriers to participation. Pharmaceutical companies have been actively exploring the potential to replace or augment traditional in-person trials with data technologies. For example, French pharmaceutical company Sanofi extended a clinical trial that had previously required participants to regularly visit the trial site in order for organizers to collect data regarding participants' weight, blood pressure, and blood glucose, by giving them connected sensors and wireless technology to record and share this data from their homes.[33] GlaxoSmithKline sponsored a study to demonstrate the feasibility of using a smartphone and app to record survey data from rheumatoid arthritis patients, as well as using the phone's accelerometer to record wrist-motion exercises, finding the accelerometer data could be much more accurate than motion-evaluation exercises performed in-person with a physician.[34] And Novartis has partnered with Apple to use Apple's ResearchKit, which helps researchers develop apps to collect and share medically relevant data, such as biometric sensor data

and user-inputted information, from smart devices to improve clinical trial recruitment and administration.[35]

Given site visits can cost between $3,000 and $7,000 per patient, and studies can involve dozens of visits and hundreds of patients, the potential for remote data collection could dramatically reduce the cost of conducting a clinical trial.[36]

New technologies such as the Internet of Things (IoT) provide opportunities to collect large amounts of data outside of a traditional health care context, known as real-world data (RWD), potentially providing valuable evidence to help inform drug evaluation, known as real-world evidence (RWE).[37] In December 2018, the FDA published the framework for its Real-World Evidence Program, which provides guidance about how to incorporate RWD into clinical trials in order to create meaningful RWE.[38]

FDA REVIEW

The FDA review process consists of evaluating data from clinical research to determine a drug's safety and efficacy. This data can be extensive, consisting of reports on all studies, and analysis from clinical and preclinical research.[39] Just as pharmaceutical researchers are using data to improve drug development and discovery, the FDA is leveraging data to help bring drugs to market faster.

In particular, the FDA has begun to conduct real-time analysis of clinical research data as soon as it becomes available, enabling it to approve or reject a drug much more quickly.[40] The FDA conducts this real-time analysis as part of its Real-Time Oncology Review (RTOR) pilot program, which focuses on bringing cancer drugs to market, and more quickly approving existing drugs for new applications.[41] In September 2018, the FDA's RTOR program approved an expanded use of the breast cancer drug Kisqali just one month after the drug's submission date.[42] This was a dramatic reduction in turnaround time, as the FDA's review process for new drugs can take anywhere from six months to two years.[43] Though the FDA's plans to use real-time analysis are so far limited to the RTOR pilot, which has a narrow scope, this approach could be adopted more broadly throughout the agency as it develops the model.

FDA POST-MARKET SAFETY MONITORING

Despite rigorous testing and evaluation in the prior two stages of the drug lifecycle, it is impossible to gain a complete understanding of a drug's safety and efficacy at the time of approval.[44] This is why, in 2007, the FDA launched its Sentinel Initiative to helped monitor the safety of drugs after they have reached the market.[45] The initiative analyzes large amounts of data from electronic health records (EHRs), insurance claims, and partners to create a post-market risk-identification system, and led to the 2016 launch of the full-scale Sentinel System that covers hundreds of millions of patients.[46] Historically, the Sentinel Initiative's goal has been to keep unsafe drugs off the market. However, in the FDA's Sentinel System Five

Year Strategy 2019–2023, the agency stated one of its goals for the future of the Sentinel System is to "explore the utility of real-world data as a tool to support drug development and assess medical product performance."[47] Specifically, the FDA intends the Sentinel System to help the agency define, test, and shape regulatory-grade data, methods, and analytical standards for the drug development process.[48] In effect, this would allow the FDA to adjust the program's mission to help better develop safer drugs with the help of data, rather than just police drug safety. For example, while the Sentinel System could help link disease registries that have longitudinal data about patient populations with other RWD sources, the registries could become more effective at engaging patients for clinical trials.[49]

WHY DATA-DRIVEN DRUG DEVELOPMENT IS HAPPENING NOW

Data has always had a role in drug development, but there are several reasons the opportunity for data-driven drug development is greater than ever.

ELECTRONIC HEALTH RECORDS

Over the past decade, the health care sector has undergone a digital transformation with the adoption of EHR systems. As of 2017, 96 percent of hospitals had adopted EHR systems certified by the Office of the National Coordinator for Health Information Technology (ONC)—as have 80 percent of physicians.[50] Widespread EHR use has substantially increased the quantity and quality of health data available to biomedical researchers, which in turn delivers crucial benefits at key stages of the drug development lifecycle.

NEW DATA SOURCES

Better data collection and analysis technologies are making it increasingly possible to take advantage of genetic data in the drug development lifecycle.

Historically, medical treatment was dictated by what worked for the average person, thereby making it rely on a general, one-size-fits-all approach.[51] But increasingly, researchers are finding that human differences play a key role in the effectiveness of treatments. The incorporation of genetic data to guide drug development allows for treatments tailored to be effective for particular groups, and even individuals.

The largest contributing factor to the feasibility of using genetic data is the increasing utility and cost-effectiveness of genetic-sequencing technologies. The Human Genome Project (HGP), a $2.7 billion initiative, created the first reference sequence of the human genome in June 2000 after 13 years of work. HGP researchers have estimated the first genome

sequencing cost between $500 million and $1 billion.[52] Since 2000, genome sequencing has become cheaper and faster by orders of magnitude, in part because of improvements in computing power. Today, modern sequencing machines can sequence a whole human genome in under an hour, for under $1,000—and the cost is expected to fall even further in coming years.[53] Genome sequencing has become so accessible governments have launched major initiatives to compile massive troves of human genetic data to accelerate medical research. In December 2018, the U.K.'s National Health Service hit its genome-sequencing target for its 100,000 Genomes Project, which is making data available to researchers developing treatments for cancers and rare diseases.[54]

Importantly, genomics can benefit drug development in more ways than by simply incorporating patient genetic data. Researchers are increasingly sequencing the genetic information of pathogens, including viruses, bacteria, fungi, and parasites.[55] For example, researchers recently sequenced the DNA of over 20,000 pneumonia strains from infected individuals in 51 countries.[56] This data revealed 621 new strains of the bacteria, and enabled researchers to identify how the bacteria population responded to various vaccines, which in turn will allow others to better develop new versions of pneumococcal vaccines.[57]

The use of genetic data is particularly promising for the development of treatments for "rare diseases"—those that affect fewer than 200,000 people.[58] While rare disease affects between 25 and 30 million Americans overall, there are approximately 7,000 rare diseases, thus making each disease's affected population potentially quite small.[59] Given the high cost of drug development, and such small patient populations, pharmaceutical companies have often viewed developing new drugs for rare diseases as not cost-effective. However, 80 percent of rare diseases are genetic, meaning advances in genome sequencing and analysis have made it more feasible than ever to spot the genetic marker of these diseases in order to help inform the development of new treatments.[60] For example, genomics company FDNA specializes in compiling genetic data, symptoms, and other information about rare diseases to make this information more easily available to researchers.[61]

Importantly, the maturation of genomics has accompanied somewhat related disciplines that could have similarly large potential impacts on drug development. Many of the technologies and techniques developed in pursuit of sequencing the human genome have spurred the development of the "omics" revolution, which includes advancements in fields such as proteomics (the study of proteomes, which are sets of proteins produced by an organism); transcriptomics (the study of an organism's RNA transcripts, which are responsible for the expression of genes); and metabolomics (the study of an organism's metabolites).[62] These advancements can make a previously untapped wealth of data available to

biopharmaceutical researchers to help them better understand human biology and guide drug development.

Additionally, new kinds of data from nontraditional sources are increasingly available to aid the drug development process. Biometric, lifestyle, and environmental data—the collection of which is made possible by the proliferation of low-cost IoT devices such as fitness trackers and air-quality monitors—are rich troves of valuable information that help researchers and regulators better evaluate a new drug's effectiveness.

The increased availability of data from a variety of sources—and the technology with which to process it—makes it feasible to pursue precision medicine, defined as "an emerging approach for disease treatment and prevention that takes into account individual variability in genes, environment, and lifestyle for each person."[63]

Combining genetic, EHR, and such nontraditional data sources as fitness trackers can enable insights that would be previously unobtainable, such as understanding how lifestyle and environmental factors influence disease risk and treatment. For example, as of July 2019, NIH lists 68 studies it funds involving the use of fitness-tracker data to study disease, such as how exercise may reduce the cardio-toxic effects of a breast cancer drug in newly diagnosed women, which could better inform evaluation of the drug's safety.[64] This is why, in 2016, NIH launched its Precision Medicine Initiative and established its All of Us Research Program, which aims to establish a cohort of 1 million people who donate data about their genetics, medical histories, lifestyles, and other factors in order to accelerate precision medicine research.[65]

ARTIFICIAL INTELLIGENCE

AI—the field of computer science devoted to creating computing-machines systems that perform operations analogous to human learning and decision-making—is a major driver of innovation in data-driven drug development. One of the most promising areas, applications of AI are interpreting unstructured data, which is information that is not organized in a predefined manner. Computer systems have been used to process structured data in health care for decades, and much of the data found in EHRs can be considered structured data, such as patient demographics and lab results. However, unstructured data is far more plentiful, constituting 80 percent of all clinical data in the United States.[66] Unstructured data in health care includes medical scans and imagery, handwritten doctors' notes, adverse event descriptions, and more. This data can contain valuable information for guiding drug development. Historically however, computer systems have not been able to easily process this information, making providers rely on humans to make sense of it, such as a radiologist interpreting an X-ray to evaluate whether a drug

is effective at shrinking a tumor, or a researcher reviewing a doctor's notes in a patient's health record to determine whether if they are a good candidate for a clinical trial. This abundance of unstructured data makes it difficult for providers to quickly distill large amounts of potentially useful medical information, thereby limiting its utility. For example, many EHR systems at clinical trial sites can only make use of structured data, causing trial operators to either use humans to convert unstructured data into structured data, which is resource-intensive and error prone, or omit it, which could waste useful data.[67]

Machine learning—a branch of AI that focuses on designing algorithms that can automatically and iteratively build analytical models from new data without explicitly programming the solution—is well-suited to analyzing unstructured data, unlocking this vast resource for a wide variety of new and valuable applications to improve drug development. Already, AI systems are matching or exceeding human-level performance in analyzing unstructured medical data, often in dramatically shorter amounts of time. For example, researchers at the University of Massachusetts have developed an AI system that can sift through unstructured EHR data and detect adverse drug events with 65.9 percent accuracy, outperforming traditional detection methods, which could help inform the post-market safety-monitoring stage of the drug development lifecycle.[68] And researchers at Google have developed a machine learning system that can, with 89 percent accuracy, identify in medical scans when breast cancer has metastasized, compared with a human accuracy rate of 73 percent—an approach that could improve analysis of a drug's effectiveness.[69] Similar examples of AI systems analyzing unstructured data to provide key medical insights are increasingly common, and show the potential for AI to give researchers more accurate and useful data faster than ever before.

AI already plays a prominent role in many aspects of data-driven drug development—and this role will only increase as the technology matures.

ACCELERATING DATA-DRIVEN DRUG DEVELOPMENT: CREATING A NATIONAL HEALTH RESEARCH DATA EXCHANGE

Progress in data-driven drug development will stagnate unless drug researchers gain access to more patient data. The broad sharing of patient data is beneficial to data-driven drug development in several key ways. First, participant-level data can be used to understand the results of trials, enabling researchers to better explicate the relationship between treatments and outcomes.[70] Second, researchers can use shared data to verify studies and identify cases of data fraud and research misconduct in the medical community.[71] Third, shared data can be combined and supplemented to support new studies and discoveries.[72]

Unfortunately, while patients are often willing to share their data to advance medical research, few are given the option of doing so. For example, a 2018 Stanford University study found that 93 percent of medical trial participants in the United States were willing to share their medical data with university scientists, and 82 percent with scientists at for-profit companies.[73] Additionally, overall, 53 percent of Americans were willing to share their health data with health care researchers, and younger Americans were significantly more likely to be willing to share their health data.[74]

However, the majority of regulations surrounding medical data focus on individuals' ability to restrict the use of their medical data, with scant attention paid to supporting the ability to share personal data for the common good. Because of this, researchers, research funders, and regulators have struggled to establish norms and practices for sharing health data.

At present, there is no simple way for patients to contribute their personal data for broad use in medical research. Individuals can consent to specific uses of specific portions of their medical data, such as sharing records associated with participation in a particular clinical trial. However, this consent does not extend to the entirety of an individual's medical data or enable the reuse of this data for similarly beneficial research.

Policymakers should address this problem by substantially expanding and increasing funding for NIH's Precision Medicine Initiative to launch a new National Health Research Data Exchange that would allow patients to quickly and easily share their data for the purpose of advancing medical research. The result would be any researcher or institution that qualifies for access to this data being able to use it for research purposes, including drug development.

Health-data registries, sometimes called patient registries, clinical data registries, or disease registries, are already a common resource in health care research and provide a significant precedent for the National Health Research Data Exchange.[75] For example, registries exist for Alopecia Areata, colon cancer, lupus, preeclampsia, sarcoidosis, some rare diseases, and many others.[76] These registries can vary in utility and provide services such as granting partnered researchers with registry participants' clinical data to advance research into understanding disease, develop new treatments, and help researchers recruit for clinical trials.[77] While these registries have enabled large amounts of valuable research, they require patients to proactively seek them out in order to participate, only focus on specific conditions, may not accumulate large enough datasets to be useful, and may not collect data in standardized formats, among other limitations.[78]

A National Health Research Data Exchange would deliver the benefits of these registries on a massive scale while eliminating many of their shortcomings. Unlike registries that compile participants' health data into a

centralized database, the National Health Research Data Exchange could instead rely on a decentralized architecture with a standardized application programming interface (API), which are protocols that enable access to the data of an application, database, or other service, to provide access to data within EHR systems. This would ensure these records are longitudinal and always up to date, rather than just a snapshot of a patient's records. EHR systems that meet ONC's meaningful use criteria are already required to have API capabilities to provide patients access to their data. Creating APIs for research would have the key benefit of ensuring qualified researchers can always access longitudinal, up-to-date patient data regardless of where it is housed.

While developing the National Health Research Data Exchange would be a significant initiative, creating such a national platform for health data could deliver massive benefits for drug development, health care research, and public health. As a first step, Congress should direct NIH, the Centers for Medicare and Medicaid Services (CMS), ONC, the Centers for Disease Control, the FDA, and other relevant health care agencies to develop a roadmap for the establishment of the Exchange.

EHR Integration

Encouraging adoption of the National Health Research Data Exchange API into EHR systems would be no small task, however the mechanism to accomplish this already exists. CMS's and ONC's Meaningful Use program from 2010 to 2017 implemented progressively higher standards for EHR systems to be eligible to be certified for use with patients covered by Medicare and Medicaid.[79] The final stage of Meaningful Use standards includes certification criteria for APIs, recognizing their utility in facilitating the sharing of health data.[80] To support this initiative, CMS developed an incentive payment program for eligible health care providers to encourage broader adoption of certified EHR technologies.[81] The Meaningful Use and incentive programs have since been renamed the Promoting Interoperability programs, which continue to encourage the development and adoption of EHR systems that emphasize the exchange of health data between health care stakeholders.[82]

Policymakers should direct CMS and ONC to require the integration of the National Health Research Data Exchange API for an EHR system to be eligible for certification and incentive payments through the Promoting Interoperability programs. Establishing the original Meaningful Use standards involved extensive deliberation with stakeholders such as hospitals and EHR developers. While expanding the program to include a specific API integration for research purposes would be a significantly smaller task by comparison, CMS and ONC should nonetheless engage with these stakeholders and carefully determine the best way to implement this requirement in order to maximize the effectiveness of the National Health Research Data Exchange.

Patient Participation

The National Health Research Data Exchange should be an opt-out system, whereby patients are enrolled in it automatically unless they choose to not participate. There are multiple reasons to make it opt-out. First, as noted, past research has shown that patients are generally willing to share their health data for research.[83] Second, opt-out nudges people to share their data, which is a socially beneficial behavior. In contrast, the rates of enrollment with an affirmative opt-in requirement would likely be much lower. Third, opt-out allows those who do not want to participate the option to not do so.

However, while patients should have the ability to opt out of the National Health Research Data Exchange, there should also be incentives to participate. First, patients should receive financial incentives. There are many options for doing this. For example, it could come in the form of modestly reduced Medicare tax payments to participants, to be offset by increased Medicare taxes for nonparticipants. Second, patients should receive nonfinancial incentives, such as prioritization over nonparticipants for inclusion in relevant clinical trials. Given the enormous boon the National Health Research Data Exchange would be to drug development, and thus future treatments for all patients, such compelling incentives are entirely justifiable. Some may consider this differential treatment of individuals to be unfair, but such an argument ignores the notion that it is more unfair for individuals to opt out of sharing their data for research purposes yet expect to benefit from data-driven drug development made possible by others' willingness to share their data.

Importantly, participants should also be able to have their data donated to the National Health Research Data Exchange after their death so it remains available to researchers as a permanent resource. A donation option would provide a valuable opportunity to incentivize participation from individuals who might prefer to opt out of sharing their health data with the National Health Research Data Exchange while they are alive but are comfortable with sharing their data after death.

Data Access

Access to patient data through the National Health Research Data Exchange should be governed by a contractual model to guard against misuse. NIH's All of Us Research Program is currently developing a research protocol to ensure broad access to its data in ways that meet researchers' needs while still protecting privacy and security.[84] This model could likely be easily adapted for researchers wishing to access data through the National Health Research Data Exchange.

Furthermore, all data accessed through the National Health Research Data Exchange would be subject to the Health Insurance Portability and Accountability Act (HIPAA). Any participating organization that manages data from the National health Research Data Exchange in a way that violates HIPAA would be subject to penalties described in HIPAA and would also face additional penalties related to their access to the exchange, such

as a temporary or permanent ban from participation, depending on the severity of the violation.

Ensuring Data Integrity Through a Unique Patient Identifier

A major obstacle to the utility of a National Health Research Data Exchange is inaccurate data. Though EHR usage is commonplace, healthcare providers do not have an accurate and efficient method of matching patients' records across different systems. Most EHRs use a technique called statistical matching to identify patient records based on attributes, such as name, date of birth, and gender, although the exact set of attributes used varies by system. However, statistical matching is unreliable and prone to error.[85] Patients may be misidentified if other patients share the same attributes, or their records may not be found if different systems store data in different formats, or certain data is missing.[86] Even when statistical-matching algorithms use existing identifiers, such as Social Security numbers, they still generate errors because many individuals do not have these identifiers, have more than one, or do not want to disclose them.[87] Because of this, it is only possible to link 90 to 95 percent of patient records uniquely between different datasets, resulting in duplicated and fragmented records.[88] A National Health Research Data Exchange built around statistical matching would not only pull in erroneous data, but would also risk incorporating data from misidentified patients that did not consent to sharing their data.

Two decades ago, HHS cited an "urgent and critical" need to create a standardized system of unique patient identifiers for health care.[89] Using unique patient identifiers would allow health-care providers to consistently and accurately link electronic health records across different systems.[90] Indeed, the original language of HIPAA called for the creation of a national universal patient-identifier system, but subsequent legislation blocked funding for implementing such a program.[91]

Congress should direct HHS to implement a unique patient identifier, as originally intended by HIPAA. This would have far-reaching benefits for the entire health care sector.[92] For the purposes of a National Health Research Data Exchange, unique patient identifiers would improve data accuracy in EHRs and ensure qualified researchers have complete access to records from enrolled patients.

Addressing Privacy Concerns

Policymakers would undoubtedly receive pushback on a National Health Research Data Exchange from people and advocacy organizations concerned about its implications for patient privacy. However, policymakers should recognize that HIPAA protections still apply and are no different than when providers use patient data for research today—the main difference would be much more data available for research.

Pharmaceutical companies, insurance companies, and other corporate entities privacy activists may believe could exploit this patient data for discriminatory or otherwise nefarious purposes already operate in a highly regulated sector and therefore already prioritize legal compliance. Given access to data through the National Health Research Data Exchange would be granted on a contractual basis to qualified researchers, misusing this data would constitute a breach of contract and potentially a breach of HIPAA, and result in significant penalties, which in addition to severe fines could include a permanent ban on access to future data, thus serving as a strong deterrent against nefarious use.

The concept of a National Health Research Data Exchange may also receive pushback from people who view it as unfair for patients to donate their data to advance drug development research that may have commercial and cost-saving applications while their health care costs remain unchanged. This concern is understandable but shortsighted, as benefits from sharing health data would come in many forms. For example, patients and their descendants would eventually benefit from new treatments that previously were impossible to create without their data. And more generally, donors and non-donors alike would benefit indirectly from corresponding lower health care costs and living in a country with healthier individuals. More broadly, donating personal data to advance medical research creates an enormously valuable social benefit with potentially life-saving implications for the entire population. Should policymakers decide to implement additional incentives for patient participation in the National Health Research Data Exchange, such incentives could be viewed as compensation for sharing data. Additionally, it is important to keep in mind this data is non-rivalrous, meaning sharing data with the National Health Research Data Exchange would not preclude patients from doing anything else with their data, including monetizing it in other capacities.

ADDITIONAL WAYS TO ACCELERATE DATA-DRIVEN DRUG DEVELOPMENT

Data-driven technologies are already improving drug development. However, this transformation can and should happen faster and more deeply. While stakeholders have strong incentives to accelerate the development and deployment of these technologies, additional barriers impede progress. These include issues related to developing AI skills, the availability of institutional and nontraditional data, outdated regulatory processes, and equity considerations. While there are many policies the United States should implement to increase its competitiveness in AI, there are specific actions policymakers can take to support AI as it relates to data-driven drug development.[93]

INCREASING THE AVAILABILITY OF INSTITUTIONAL AND NONTRADITIONAL DATA

As data becomes more important for drug development, policymakers should recognize where barriers to data sharing exist. In some cases, regulatory obstacles are to blame. For example, as Jessica Jardine Wilkes has noted in the *BYU Law Review*, due to HIPAA's complexity and the severe penalties for noncompliance, "many covered entities ... 'have implemented business practices in the name of privacy and security that have no basis in law'" in order to hedge their risk of potentially running afoul of HIPAA, resulting in widespread reluctance to share health data.[94] In others, cultural issues are the culprit. For example, researchers themselves have a proprietary interest in data they produce, while academic researchers seeking to maximize publications may guard data.[95] Despite broad recognition from funding bodies, policymakers and researchers that data sharing would be beneficial, academic researchers rarely make their data available for others. The NIH, for example, requires a data-sharing plan for big-ticket funding, but recognizes that proprietary interests may impede that sharing.[96]

Policymakers should pursue an array of opportunities to increase the sharing of data by institutions traditionally involved in drug development as well as from nontraditional sources.

Expanding Access to Institutional Data

Institutions involved in the drug development lifecycle that rely on patient data, such as pharmaceutical companies, research institutes, and government agencies, aggregate this information and use it to conduct research, improve clinical trials, and develop a wide variety of other valuable data products. Increasing the availability and sharing of institutional data is just as important as increasing the availability of patient data for accelerating data-driven drug discovery.

In many cases, researchers, pharmaceutical companies, and health agencies may be willing to share data with each other to advance data-driven drug development, but are unable to do so. These stakeholders could all benefit substantially from sharing data to develop new drugs, reduce costs, and gain insights, but lack mechanisms to do so in ways that respect the proprietary and sensitive nature of this data. The United Kingdom is working to address this challenge by developing a model for data trusts, which it defines as "not a legal entity or institution, but rather a set of relationships underpinned by a repeatable framework, compliant with parties' obligations to share data in a fair, safe, and equitable way."[97] Without a coordinating body such as a government agency specifically devoted to developing and supporting these models, organizations may be slow to develop them on their own. Indeed, while stakeholders have made some progress in developing data-sharing agreements to advance drug development, they typically involve only a handful of large pharmaceutical firms or focus on specific medical conditions.[98]

HHS should adapt and pilot this model for data trusts to facilitate data sharing among academia, businesses, and government institutions involved with data-driven drug development.

Another key opportunity to increase the amount of institutional data available to guide drug development is to better enforce the publication of clinical trial results. This is particularly useful for increasing the transparency around the potential risks of drugs, as well as for allowing other stakeholders to learn what does not work so they can avoid redundant research. As Srini Ramanathan, vice president of development sciences at Horizon Pharma, put it, "A lot of companies have giant databanks of compounds that they've tested that worked and didn't work … we learn by how we succeed and how we fail."[99] While some researchers share their clinical trial results in medical journals, a large share of this data is never publicized. A 2012 study found that between 25 and 50 percent of all clinical trials are never published or are only published years after the fact.[100] While it is understandable a company may want to avoid calling attention to its unsuccessful research, the study found that nearly half of all clinical trials funded by NIH went unpublished for 30 months after completion, and 32 percent were never published.[101]

What makes this particularly concerning is Congress attempted to solve this problem with the FDA Amendments Act of 2007, requiring the public reporting of clinical trials, with some exceptions, with strict penalties for noncompliance.[102] However, a 2015 investigation from medical-journalism publication *STAT* found that "Most research institutions—including leading universities and hospitals in addition to drug companies—routinely break a law that requires them to report the results of human studies of new treatments to the federal government's ClinicalTrials.gov database."[103] Of the 9,000 studies examined, 74 percent of industry trials were either reported late or not at all, while 90 percent of trials run by academic institutions were reported late or not at all.[104] This violation was likely due to, at the time of the study, neither the FDA nor NIH ever having enforced a penalty for noncompliance, which could be as high as $10,000 for each day a trial goes unreported beyond the deadline.[105]

Fortunately, clinical-trial reporting has increased substantially since 2015, likely due in large part to the *STAT* investigation: 72 percent of all trial results were disclosed as of September 2017, which the sharpest increase in adherence coming from academic and nonprofit research institutes.[106] However, shortcomings still remain, as 2 in 5 trials are reported after the deadline has passed, and certain institutions still have woefully inadequate reporting rates.[107] The FDA's finalized rule for penalizing noncompliance went into effect in January 2018, but neither the FDA nor NIH have expanded their enforcement staff.[108] Congress should ensure these agencies have the necessary resources to fully enforce the reporting law, and direct the agencies to be more aggressive about penalizing noncompliance to ensure this valuable data is widely available.

Finally, policymakers should look for opportunities to increase data sharing with international partners to advance data-driven drug development. The FDA's Center for Devices and Radiological Health, for example, has an information exchange program with other national regulators to exchange regulatory data of mutual interest, allowing them to share resources to make more informed decisions.[109] However, any such partnerships should be based on mutual benefit and sharing, with the government only entering into partnerships with nations that engage in mutual levels of sharing and adequately protect biopharmaceutical intellectual property.

Expanding Access to Nontraditional Data

The FDA's Real-World Evidence Program is encouraging as it signals that regulators are both aware of the value of new kinds of data sources for aiding drug evaluation and the potential challenges in using this data. However, the framework provides little guidance about newer, nontraditional data technologies, simply stating, "FDA will explore strategies for filling gaps in data that may be difficult to obtain from currently used EHRs and medical claims data, including exploring the use of mobile technologies, electronic patient reported outcome tools, wearables, and biosensors."[110] The FDA should prioritize the development of this guidance as quickly as possible, and update it regularly to account for new technologies.

MODERNIZING REGULATORY PROCESSES

Medicine is a highly regulated field, and as such, the ability for regulators to take advantage of data and respond to new data technologies will greatly influence the growth of data-driven drug development.

The FDA's Sentinel System shows great promise to transform the role of the regulatory review process for the better, and the FDA should be commended for it. Congress should ensure the FDA has the resources necessary to achieve this vision for the Sentinel System.

Additionally, given the ability of new technologies to make it easier to share data and conduct more decentralized clinical trials, it is increasingly feasible, and sometimes desirable when considering participant diversity, to conduct global trials across national lines. Over the past several decades foreign clinical trials (FCTs) have become increasingly common due to factors such as lower costs, less difficulty in recruiting patients, and efforts from other countries to attract research and clinical trials.[111] The FDA accepts FCT data, provided the trials were conducted in accordance with FDA standards. However, limited resources and the increase in FCTs mean the agency can only inspect a limited number of foreign sites. In 2008, the FDA inspected just 0.7 percent of foreign clinical sites, compared with 1.9 percent of domestic sites.[112] Given the potential value of this data for drug development, Congress should encourage the FDA and provide the resources necessary to increase FCT inspections, harmonize regulatory standards across national lines to meet the FDA's satisfaction, and adopt risk-assessment analytics tools to prioritize inspections for high-

risk sites.[113] Again, such cooperation should only be granted to nations that extend similar cooperation to the United States.

PROMOTING EQUITY

Despite the potential of data-driven drug development, its benefits mean less to people who are not represented in the data used to drive new discoveries and evaluate new treatments. This is a symptom of data poverty—the social and economic inequalities that arise from a lack of collection or use of data about individuals or communities.[114] Data poverty is a particular challenge for data-driven medicine as a whole. Historically, racial and ethnic minorities, as well as women, have been underrepresented in clinical trials.[115] For example, Hispanics represent approximately 16 percent of the U.S. population, but only 1 percent of clinical trial participants.[116] When certain groups are underrepresented in the data, the decisions made about the safety and efficacy of treatments for patients may be biased. For example, women are more likely to have an adverse reaction to a drug and may respond differently to medical devices. Some drugs have even been taken off the market because of effects to women that were missed in clinical trials. Similarly, studies have found that various racial and ethnic groups respond differently to certain medications.[117]

Data poverty is not just limited to clinical trials. The majority of participants in genomics research are of European descent, which limits the utility of precision-medicine treatments. As Jacquelyn Taylor, a health-equity researcher at New York University, explained, "It's hard to tailor treatments for people's unique needs, if the people who are suffering from those diseases aren't included in the studies."[118]

Policymakers have options to help address these equity concerns. The FDA began investigating how well demographic subgroups are represented in clinical trials, and whether subgroup-specific safety and effectiveness data became available after the passage of the 2012 Food and Drug Administration Safety and Innovation Act.[119] And in 2015, the agency began publishing "Drug Trial Snapshots" illustrating the diversity of clinical trial participants for new drugs.[120] However, the FDA has not established diversity guidelines for clinical trials.[121] There is some justification for this, as certain drugs are intended only for narrow demographic subgroups, which means there is little need to include participants for clinical trials that reflect the broader population. Though the FDA lacks authority to require specific levels of diversity in clinical trials, it does attempt to encourage diversity in other capacities, such as by developing educational materials for patients about clinical trial participation and best practices for recruiting women for trials.[122] The FDA should expand on these efforts and engage with pharmaceutical companies to raise awareness of the importance of diversity.

Additionally, NIH's All of Us Research Program will go a long way toward increasing the diversity of genetic data available to researchers. Policymakers should nevertheless support expansion of the program.

The Obama administration budgeted $215 million for the Precision Medicine Initiative to develop its million-person research cohort at launch.[123] By contrast, China launched its 15-year Precision Medicine Initiative in 2018 with $9.2 billion in funding with the goal of sequencing 100 million genomes. [124]

Policymakers ensure the All of Us Research Program can continue to ramp up operations to begin developing its cohort, but also develop near- and long-term goals to grow the program, including expanding the cohort beyond 1 million people, with a focus on representation that accurately reflects the diversity of the United States. To help accomplish this, Congress should appropriate additional funding so NIH can hire the workforce and develop the infrastructure to steadily grow this program year after year.[125]

Importantly, the National Health Research Data Exchange would go a long way toward promoting equity by providing an opportunity for every person in the United States to broadly share their health data for research, rather than just individuals who are particularly motivated to do so or are easy for researchers to target.

DEVELOPING AI SKILLS

AI is underlying many valuable developments in data-driven drug development. However, the United States is struggling to meet the demand for workers with AI skills, which limits progress in all applications of AI.[126] AI could change the skills involved in drug discovery. For example, it can automate much of the HTS process, making the skills to do so manually less useful while increasing the demand for computational skills necessary to develop and use AI.[127]

While some AI skills will be unique to the biopharmaceutical industry, many will not. And to the extent companies compete for scarce AI talent, that shortage will slow AI-driven drug development. There are a number of steps policymakers should take. State and federal lawmakers should promote data-science education, especially at and beyond high school, in order to raise the baseline computational skills of workers. This could include all states that allow computer science and statistics courses to count for math or science graduation requirements; prioritizing making computer science and statistics courses available in every high school; increasing the number of qualified computer science teachers; and doubling the number of science-, technology-, engineering-, and mathematics-focused charter schools.[128]

At the higher-education level, policymakers should create compelling incentives for computer science education. The National Science Foundation should provide grants to schools for implementing programs to increase computer science enrollment and retention.[129] The federal government should also require increased transparency as a prerequisite for certain educational funds. For example, schools should be required to monitor and disclose the number of computer science applicants,

prospective majors, and retention rates in order to be eligible for certain federal benefits.[130]

Additionally, Congress should enable more foreign AI talent to live and work in the United States. While attracting foreign-born AI talent is useful for U.S. competitiveness in AI broadly, the pharmaceutical industry in particular relies heavily on foreign-born workers. A 2014 study found that while immigrants made up 13 percent of the U.S. population in 2011, they made up 17 percent of pharmaceutical industry workers.[131] And by some estimates based on 2015 data, immigrants constitute 23 percent of the pharmaceutical workforce today.[132] The importance of foreign-born talent in drug discovery in particular is even more evident, making up one-third of the research and development industry.[133]

Policymakers should make it easier for foreign AI talent to help advance drug development in the United States in two ways. First, though the biopharmaceutical industry on average does not employ a large number of workers on high skilled, H-1B visas, policymakers should nonetheless increase the cap on H-1B visas to ensure U.S. pharmaceutical firms can more easily access AI. Additionally, policymakers should exempt international students with science, technology, engineering, and mathematics (STEM) degrees with job offers from Green Card caps that restrict how foreign born, U.S.-educated graduates from each country are permitted to work in the United States. A group of senators has recently introduced legislation to do this, titled the Keep STEM Talent Act of 2019.[134]

CONCLUSION

Data-driven innovation promises to be even more transformative in medicine than in many other sectors. Given the uniquely high stakes for data-driven drug development, and its potential to save lives and improve quality of life, policymakers need to push for policies to accelerate the development and deployment of data technologies, increase the accessibility of valuable data—particularly through the development of a National Health Research Data Exchange—modernize regulatory processes around the potential of data, and ensure the benefits of data-driven drug development flow to all.

REFERENCES

1. Note: the FDA considers the drug development lifecycle to have five stages, with the "preclinical research" coming before the clinical research stage. For the sake of clarify, this report only examines the four listed stages, in which the benefits of data-driven medicine are particularly evident.
2. Glenn E. Croston, "The Utility of Target-Based Discovery," *Expert Opinion on Drug Discovery 12*, no. 5 (2017): 427–429, https://tandfonline.com/doi/full/10.1080/17460441.2017.1308351.
3. Daniel P. Russo and Hao Zhu, "Accessing the High Throughput Screening Data Landscape," *Methods in Molecular Biology*, (2017), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5613246/.
4. A first-in-class drug is the first drug of its kind, whereas a best-in-class drug is the most effective drug of a class that has already been demonstrated to be effective. J. Eder, R. Sedrani, and C. Wiesmann, "The Discovery of First-In-Class Drugs: Origins and Evolution," *Nature Reviews Drug Discovery* (August 2014), 577–587, https://www.ncbi.nlm.nih.gov/pubmed/25033734?dopt=Abstract.
5. Joanne Kotz, "Phenotypic Screening, Take Two," *Science-Business eXchange 5*, no. 15 (2012): 380, https://www.nature.com/scibx/journal/v5/n15/full/scibx.2012.380.html.
6. Joanne Owens, "Phenotypic Versus Target-Based Screening for Drug Discovery," *Technology Networks*, April 24, 2018, https://www.technologynetworks.com/drug-discovery/articles/phenotypic-versus-target-based-screening-for-drug-discovery-300037; Bridget K. Wagner, "The Resurgence of Phenotypic Screening in Drug Discovery and Development," *Expert Opinion on Drug Discovery 11*, no. 2 (2016): 121–125, https://www.tandfonline.com/doi/full/10.1517/17460441.2016.1122589.
7. WF. An, "Fluorescence-Based Assays," *Methods in Molecular Biology* (2009), 97–107, https://www.ncbi.nlm.nih.gov/pubmed/19347618.
8. Ruairi J. Mackenzie, "How AI Can Speed Up High Content Screening Analysis," *Technology Networks*, November 14, 2018, https://www.technologynetworks.com/informatics/blog/how-ai-can-speed-up-high-content-imaging-analysis-311841.
9. Ibid.
10. Christian Scheeder, Florian Heigwer, and Michael Boutros, "Machine Learning and Image-Based Profiling in Drug Discovery," *Current Opinion in Systems Biology 11* (August 2018): 43–52, https://www.sciencedirect.com/science/article/pii/S2452310018300027.
11. J. Patel, "Science of the Science, Drug Discovery and Artificial Neural Networks," *Current Drug Discovery Technology 10*, no. 1 (2013): 2–7, https://www.ncbi.nlm.nih.gov/pubmed/22725688.
12. Robert D. Atkinson, "Drug Price Controls Will Be More Pain than Gain," *The Hill*, November 10, 2018, https://thehill.com/opinion/healthcare/416068-drug-price-controls-will-be-more-pain-than-gain; "Cost to Develop and Win marketing Approval for a New Drug is $2.6 Billion," Tufts Center for the Study of Drug Development, News Release, November 18, 2014,

https://static1.squarespace.com/static/5a9eb0c8e2ccd1158288d8dc/t/5ac66adc758d46b001a996d6/1522952924498/pr-coststudy.pdf.

13. Penn State News, "Large-Scale Study Reveals new Candidates for Malaria Prevention Drug," *Penn State News*, news release, December 6, 2018, https://news.psu.edu/story/550770/2018/12/06/research/large-scale-study-reveals-new-candidates-malaria-prevention-drug.
14. Sean Ekins et al., "Combining Metabolite-based Pharmacophores with Bayesian Machine Learning Models for Combining Metabolite-Based Pharmacophores for *Mycobacterium Tuberculosis* Drug Discovery," *PLOS One* (October 2015), https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0141076.
15. Ibid.
16. "Structure-Guided Drug Discoveries," The Petsko & Ringe Laboratories at Brandeis University, accessed July 22, 2019, http://www.bio.brandeis.edu/prLab/drug.html.
17. "Fully Automated Facility for Drug Discovery To Be Built in UK," *Pharmaceutical Technology*, June 7, 2018, https://www.pharmaceutical-technology.com/news/fully-automated-facility-drug-discovery-built-uk/.
18. Bethany Halford, "How DNA-Encoded Libraries Are Revolutionizing Drug Discovery," *Chemical & Engineering News*, June 19, 2017, https://cen.acs.org/articles/95/i25/DNA-encoded-libraries-revolutionizing-drug.html.
19. Ibid.
20. Ibid.
21. Frank Vinluan, "Relay Raises $400M to Kick protein Motion Drug R&D Into High Gear," *Xconomy*, December 20, 2018, https://xconomy.com/boston/2018/12/20/relay-raises-400m-to-kick-protein-motion-drug-rd-into-high-gear/; Frank Vinluan, "Relay Tx Raises $63M to Advance Protein Motion Drugs Into the Clinic," *Xconomy*, December 14, 2017, https://xconomy.com/boston/2017/12/14/relay-tx-raises-63m-to-advance-protein-motion-drugs-into-the-clinic/.
22. Johns Hopkins Bloomberg School of Public Health, "Cost of Clinical Trials for New Drug FDA Approval Are Fraction of Total Tab," news release, September 24, 2018, https://www.jhsph.edu/news/news-releases/2018/cost-of-clinical-trials-for-new-drug-FDA-approval-are-fraction-of-total-tab.html.
23. Ibid; Aylin Setkaya et al., "Examination of Clinical Trial Costs and Barriers for Drug Development," (submitted to the U.S. Department of Health and Human Services Office of the Assistant Secretary for Planning and Evaluation, July 2014), https://aspe.hhs.gov/system/files/pdf/77166/rpt_erg.pdf.
24. Jack Kaufman, "The Innovative Startups Improving Clinical Trial Recruitment, Enrollment, Retention, and Design," *MobiHealthNews*, November 30, 2018, https://www.mobihealthnews.com/content/innovative-startups-improving-clinical-trial-recruitment-enrollment-retention-and-design; "The Next Step: Using Ai to Formulate Clinical Trial Research Questions," Anju Life Sciences Software, accessed July 22, 2019, https://anjusoftware.com/about/all-news/insights/ai-trial-research-questions/.

25. Jack Kaufman, "The Innovative Startups Improving Clinical Trial Recruitment, Enrollment, Retention, and Design," *MobiHealthNews*, November 30, 2018, https://www.mobihealthnews.com/content/innovative-startups-improving-clinical-trial-recruitment-enrollment-retention-and-design.

26. Kumba Sennaar, "AI and Machine Learning for Clinical Trials – Examining 3 Current Approaches," *Emerj*, March 5, 2019, https://emerj.com/ai-sector-overviews/ai-machine-learning-clinical-trials-examining-x-current-applications/.

27. Ibid.

28. Gregory Yuaney and Pratik Shah, "Reinforcement Learning with Action-Derived Rewards for Chemotherapy and Clinical Trial Dosing Regimen Selection," *Proceedings of the 3rd Machine Learning for Health Care Conference 85* (2018): 161–226, http://proceedings.mlr.press/v85/yauney18a.html.

29. Ibid.

30. Stefan Harrer et al., "Artificial intelligence for Clinical Trial Design," *Trends in Pharmacological Sciences 40*, Issue 8 (2019): 577–591, https://www.cell.com/trends/pharmacological-sciences/fulltext/S0165-6147(19)30130-0?_returnURL=https%3A%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS0165614719301300%3Fshowall%3Dtrue#f0010.

31. Barbara Mantel-Undark, "The Search for New Drugs is Coming To Your House," *Fast Company*, August 30, 2018, https://www.fastcompany.com/90229910/virtual-clinical-trials-are-bringing-drug-development-home.

32. Ibid.

33. Ibid.

34. Ibid.

35. Andrew McConaghie, "Novartis and Apple to Scale Up Clinical Trial Collaboration," *PharmaPhorum*, January 24, 2018, https://pharmaphorum.com/news/researchkit-novartis-apple-scale-clinical-trial-collaboration/.

36. Barbara Mantel-Undark, "The Search for New Drugs is Coming To Your House," *Fast Company*, August 30, 2018, https://www.fastcompany.com/90229910/virtual-clinical-trials-are-bringing-drug-development-home.

37. *Framework for FDA's Real-World Evidence Program*, (Washington, D.C.: U.S. Food and Drug Administration, December 2018), https://www.fda.gov/media/120060/download.

38. Ibid.

39. "Step 4: FDA Drug Review," U.S. Food and Drug Administration, accessed July 22, 2019, https://www.fda.gov/ForPatients/Approvals/Drugs/ucm405570.htm.

40. "Real-Time Review of Drug Applications is New a Reality," U.S. Good and Drug Administration, September 20, 2018, https://www.fda.gov/Drugs/DevelopmentApprovalProcess/SmallBusinessAssistance/ucm621194.htm.

41. Ibid.
42. Ibid.
43. U.S. Food and Drug Administration (FDA), *PDUFA Reauthorization Performance Goals and Procedures Fiscal Years 2018 Through 2022* (Washington, D.C.: FDA), accessed July 22, 2019, https://www.fda.gov/downloads/forindustry/userfees/prescriptiondruguserfee/ucm511438.pdf; "The Drug Development and Approval process," FDA Review, accessed July 22, 2019, http://www.fdareview.org/issues/the-drug-development-and-approval-process/.
44. "Step 5: FDA Post-Market Drug Safety Monitoring," U.S. Food and Drug Administration, accessed July 22, 2019, https://www.fda.gov/ForPatients/Approvals/Drugs/ucm405579.htm.
45. "FDA's Sentinel Initiative – Background," U.S. Food and Drug Administration, accessed July 22, 2019, https://www.fda.gov/safety/fdas-sentinel-initiative/fdas-sentinel-initiative-background.
46. U.S. Food and Drug Administration, *Sentinel System Five-Year Strategy 2019–2023*, (Washington, D.C.: January 2019), https://www.fda.gov/media/120333/download.
47. Ibid.
48. Ibid.
49. Ibid.
50. "Electronic Health Record Adoption," Office of the National Coordinator for Health Information Technology, accessed July 22, 2019," https://dashboard.healthit.gov/apps/health-information-technology-data-summaries.php?state=National&cat9=all+data&cat1=ehr+adoption#summary-data.
51. "What Is Precision Medicine?" U.S. National Library of Medicine," accessed July 22, 2019, https://ghr.nlm.nih.gov/primer/precisionmedicine/definition.
52. "The Cost of Sequencing a Human Genome," National human Genome Research Institute, accessed July 22, 2019, https://www.genome.gov/27565109/the-cost-of-sequencing-a-human-genome/.
53. "DNA Sequencing Costs: Data," National Human Genome Research Institute, accessed July 22, 2019, https://www.genome.gov/27541954/dna-sequencing-costs-data/; Bradley J. Fikes, "Focus: new Machines Can Sequence Human Genome in One Hour, Illumina Announces," *The San Diego Union-Tribune*, January 9, 2017, https://www.sandiegouniontribune.com/business/biotech/sd-me-illumina-novaseq-20170109-story.html; Megan Molteni, "Now You Can Sequence Your Whole Genome for Just $200," *Wired*, November 11, 2019, https://www.wired.com/story/whole-genome-sequencing-cost-200-dollars/.
54. "UK DNA Project Hits Major Milestone with 100,000 Genomes Sequenced," *NewScientist*, December 5, 2018, https://www.newscientist.com/article/2187499-uk-dna-project-hits-major-milestone-with-100000-genomes-sequenced/.
55. Marta Gwinn et al., "Next-Generation Sequencing of Infectious Pathogens," *JAMA Insights| Genomics and Precision Health* (February 2019), https://jamanetwork.com/journals/jama/fullarticle/2725407.

56. "Pneumonia Mapped in Global Genomic Survey of Disease-Causing Bacterium," *Technology Networks*, June 11, 2019, https://www.technologynetworks.com/genomics/news/pneumonia-mapped-in-global-genomic-survey-of-disease-causing-bacterium-320499.

57. Ibid.

58. "FAQ's About Rare Diseases," National Center for Advancing Translational Sciences, accessed July 22, 2019, https://rarediseases.info.nih.gov/diseases/pages/31/faqs-about-rare-diseases.

59. Ibid.

60. "Rare Disease By The Numbers," America's Biopharmaceutical Companies, accessed July 29, 2019, https://innovation.org/about-us/commitment/research-discovery/rare-disease-numbers.

61. "About FDNA," FDNA, accessed July 29, 2019, https://www.fdna.com/about-us/.

62. E. Mathé et al., "The Omics Revolution Continues: The Maturation of High-Throughput Biological Data Sources," *Yearbook of Medical Informatics*, no. 1, (August 2018): 211–222, https://www.ncbi.nlm.nih.gov/pubmed/30157526.

63. P. Garrado et al., "Proposal for the creation of a national strategy for precision medicine in cancer: a position statement of SEOM, SEAP, and SEFH," *Clinical & Translational Oncology*, no. 4, (April 2018): 443–447, https://www.ncbi.nlm.nih.gov/pubmed/28861725.

64. "Fitness tracker," U.S. National Library of Medicine, accessed July 26, 2019, https://clinicaltrials.gov/ct2/results?cond=&term=fitness+tracker&cntry=&state=&city=&dist=; "Impact of Exercise on Mitigating the Cardio-toxic Effects of Adriamycin Among Women Newly Diagnosed With Breast Cancer," U.S. National Library of Medicine, January 3, 2019, https://clinicaltrials.gov/ct2/show/NCT03027063?term=fitness+tracker&draw=4&rank=34.

65. "About the *All of Us* Research Program," National Institutes of Health, accessed July 26, 2019, https://allofus.nih.gov/about/about-all-us-research-program.

66. "Accelerating Clinical trials Through Access to Real-World Patient Data" (InterSystems), accessed July 26, 2019, https://www.intersystems.com/isc-resources/wp-content/uploads/sites/24/Accelerating_Clinical_Trials_Through-_Access_to_Real-World_Patient_Data_WP.pdf.

67. Center for Data Innovation, "U.S. Data Innovation Day 2018: The Future of Data-Driven Medicine," YouTube, September 13, 2018, https://www.youtube.com/watch?v=UqBSeqr_7z8.

68. Fei Li, *Extraction of Information Related to Adverse Drug Events from Electronic Health Record Notes: Design of an End-to-End Model Based on Deep Learning* 6, no. 4 (2018), https://medinform.jmir.org/2018/4/e12159/; Jessica Kent, "Deep Learning Spots Adverse Drug Event sin Unstructured EHR Data," *Health IT Analytics*, November 29, 2018, https://healthitanalytics.com/news/deep-learning-spots-adverse-drug-events-in-unstructured-ehr-data.

69. Yun Liu et al., “Detecting Cancer Metastases on Gigapixel Pathology Images,” March 8, 2017, https://arxiv.org/abs/1703.02442.

70. “Sharing Clinical Research Data: Workshop Summary” (Institute of Medicine, Washington, D.C., 2013), https://doi.org/10.17226/18267.

71. David Cyranoski, “Retraction Record Rocks Community,” *Nature*, September 19, 2012, https://www.nature.com/news/retraction-record-rocks-community-1.11434.

72. “Sharing Clinical Research Data: Workshop Summary” (Institute of Medicine, Washington, D.C., 2013), https://doi.org/10.17226/18267.

73. Michelle Mello, Van Lieou, and Steven Goodman, “Clinical Trial Participants’ Views of the Risks and Benefits of Data Sharing,” *The New England Journal of Medicine*, June 7, 2018, https://www.nejm.org/doi/full/10.1056/NEJMsa1713258; Eleni Manis, “Americans Want to Share Their Medical Data. So Why Can't They?” RealClear Health, July 26, 2018, https://www.realclearhealth.com/articles/2018/07/26/americans_want_to_share_their_medical_data_so_why_cant_they_110807.html.

74. Scott Hensley, “Poll: Most Americans Would Share Health Data For Research,” *NPR*, January 9, 2018, https://www.npr.org/sections/health-shots/2015/01/09/375621393/poll-most-americans-would-share-health-data-for-research.

75. T.A. Workman, *Engaging Patients in Information Sharing and Data Collection: The Role of Patient-Powered Registries and Research Networks* (Rockville, MD: Agency for Healthcare Research and Quality, September 2013), https://www.ncbi.nlm.nih.gov/books/NBK164514/.

76. “List of Registries,” National Institute of Health, accessed July 26, 2019, https://www.nih.gov/health-information/nih-clinical-research-trials-you/list-registries.

77. T.A. Workman, *Engaging Patients in Information Sharing and Data Collection: The Role of Patient-Powered Registries and Research Networks* (Rockville, MD: Agency for Healthcare Research and Quality, September 2013), https://www.ncbi.nlm.nih.gov/books/NBK164514/.

78. Ibid.

79. “Public Health and Promoting Interoperability Programs (formerly, known as Electronic Health Records Meaningful Use),” https://www.cdc.gov/ehrmeaningfuluse/introduction.html.

80. “About APIs,” Office of the National Coordinator for Health Information Technology,” accessed July 29, 2019, https://www.healthit.gov/api-education-module/story_content/external_files/hhs_transcript_module.pdf.

81. “Public Health and Promoting Interoperability Programs (formerly, known as Electronic Health Records Meaningful Use),” https://www.cdc.gov/ehrmeaningfuluse/introduction.html.

82. “CMS Finalizes Changes to Interoperability Initiatives and EHR Incentive Program for Hospitals,” *HIMMS*, August 3, 2018, https://www.himss.org/news/cms-finalizes-changes-interoperability-initiatives-and-ehr-incentive-program-hospitals.

83. Michelle Mello, Van Lieou, and Steven Goodman, “Clinical trial Participants’ Views of the Risk and Benefits of Data Sharing,” *New England Journal of*

Medicine, no. 378 (June 2018): 2202-2211, https://www.nejm.org/doi/full/10.1056/NEJMsa1713258

84. "Researchers as Partners," National Institutes of Health, accessed July 26, 2019, https://www.researchallofus.org/about/researchers-as-partners/.

85. Bonnie Darves, "Quest for a Unique Identifier Stalled," iHealthBeat, April 8, 2014, http://www.ihealthbeat.org/insight/2014/quest-for-aunique-patient-identifier-stalled.

86. Genevieve Morris et al., "Patient Identification and Matching Final Report" (Washington, D.C.: Office of the National Coordinator for Health Information Technology, February 7, 2014), http://www.healthit.gov/sites/default/files/patient_identification_matching_final_report.pdf.

87. Charles Cooper, "For 20 Million Americans, One Social Security Number's Not Enough," *CBS News*, August 16, 2010, http://www.cbsnews.com/news/for-20-million-americans-one-socialsecurity-numbers-not-enough/.

88. Harpreet S. Sood et al., "Has the Time Come for a Unique Patient Identifier for the U.S.?" *NEJM Catalyst*, February 21, 2018, https://catalyst.nejm.org/time-unique-patient-identifiers-us/.

89. U.S. Department of Health and Human Services (HHS), "Analysis of Unique Patient Identifier Options" (Washington, D.C.: HHS, November 24, 1997), https://www.ncvhs.hhs.gov/wpcontent/uploads/2014/08/APPAVU-508.pdf.

90. Richard Hillestad et al., "Identity Crisis: An Examination of the Costs and Benefits of a Unique Patient Identifier for the U.S. Health Care System" (RAND Corporation, 2008), http://www.rand.org/content/dam/rand/pubs/monographs/2008/RAND_M G753.pdf.

91. Neil Versel, "National Patient ID System: Debate Stoked," *InformationWeek*, March 29, 2013, http://www.informationweek.com/administration-systems/nationalpatient-id-system-debate-stoked/d/d-id/1109314.

92. Daniel Castro, Joshua New, and Matt Beckwith, "10 Steps Congress Can Take to Accelerate Data Innovation" (Center for Data Innovation, May 2017), http://www2.datainnovation.org/2017-data-innovation-agenda.pdf.

93. Joshua New, "Why The United States Needs a National Artificial Intelligence Strategy and What It Should Look Like" (Center for Data Innovation, December 2018), http://www2.datainnovation.org/2018-national-ai-strategy.pdf.

94. Jessica Wilkes, "The Creation of HIPPAA Culture: Prioritizing Privacy Paranoia Over Patient Care," *BYU Law Review 2014*, Issue 5 (November 2014): 1,212–1,250, https://pdfs.semanticscholar.org/155f/5432976118eea9837deec7d01954bd23700d.pdf; Linda Dimitropoulos, "Privacy and Security Solutions for Interoperable Health Information Exchange," (Washington, D.C.: Office of the National Coordinator for Health Information Technology, December 20, 2007), https://www.rti.org/sites/default/files/resources/phase2_impactanaly.pdf.

95. "Sharing Clinical Research Data: Workshop Summary" (Institute of Medicine, Washington, D.C., 2013), https://doi.org/10.17226/18267; Benedikt Fecher, Sascha Friesike, and Marcel Hebing, "What Drives

Academic Data Sharing?" *PLoS One 10,* no. 2, (2015), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4340811/.

96. "NIH Data Sharing Policy," National Institutes of Health, accessed July 26, 2019, https://grants.nih.gov/grants/policy/data_sharing/data_sharing_brochure.pdf.

97. Dame Wendy Hall and Jérôme Pesenti, "Growing the Artificial Intelligence Industry in the UK" (London: Independent Report, October 2017), https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/652097/Growing_the_artificial_intelligence_industry_in_the_UK.pdf.

98. Stefanie Koperniak, "Applying Machine Learning to Challenges in the Pharmaceutical Industry," *MIT News*, May 17, 2018, http://news.mit.edu/2018/applying-machine-learning-to-challenges-in-pharmaceutical-industry-0517; Patricia Van Arnum, "Pharma Companies Form a Clinical Information-Sharing Consortium," *PharmTech*, June 17, 2010, http://www.pharmtech.com/pharma-companies-form-clinical-information-sharing-consortium.

99. Center for Data Innovation, "U.S. Data Innovation Day 2018: The Future of Data-Driven Medicine," YouTube, September 13, 2018, https://www.youtube.com/watch?v=UqBSeqr_7z8.

100. Joseph S. Ross et al., "Publication of NIH Funded Trials Registered in ClinicalTrials.gov; Cross Sectional Analysis," *BMJ* (2012), https://www.bmj.com/content/344/bmj.d7292.

101. Stephanie Wykstra, "A Surprising Amount of Medical Research Isn't Made Public. That's Dangerous," *Vox*, August 1, 2017, https://www.vox.com/the-big-idea/2017/8/1/16012946/clinical-trial-research-public-transparency; Joseph S. Ross et al., "Publication of NIH Funded Trials Registered in ClinicalTrials.gov; Cross Sectional Analysis," *BMJ* (2012), https://www.bmj.com/content/344/bmj.d7292.

102. Stephanie Wykstra, "A Surprising Amount of Medical Research Isn't Made Public. That's Dangerous," *Vox*, August 1, 2017, https://www.vox.com/the-big-idea/2017/8/1/16012946/clinical-trial-research-public-transparency; "Food and Drug Administration Amendments Act (FDAAA) of 2007," Food and Drug Administration, accessed July 26, 2019, https://www.fda.gov/RegulatoryInformation/LawsEnforcedbyFDA/SignificantAmendmentstotheFDCAct/FoodandDrugAdministrationAmendmentsActof2007/default.htm.

103. Charles Piller, "Failure to Report: A STAT Investigation of Clinical Trials Reporting," *STAT*, December 13, 2015, https://www.statnews.com/2015/12/13/clinical-trials-investigation/.

104. Ibid.

105. Ibid.

106. Charles Piller and Talia Bronshtein, "Faced with Public Pressure, Research Institutions Step Up Reporting of Clinical Trial Results," *STAT*, January 9, 2018, https://www.statnews.com/2018/01/09/clinical-trials-reporting-nih/.

107. Ibid.

108. Ibid.

109. "Information Exchange with Other Regulators," Food and Drug Administration, accessed July 26, 2019, https://www.fda.gov/MedicalDevices/InternationalPrograms/InformationExchangewithOtherRegulators/index.htm.

110. *Framework for FDA's Real-World Evidence Program*, (Washington, D.C.: U.S. Food and Drug Administration, December 2018), https://www.fda.gov/media/120060/download.

111. Kassa Ayalew, "FDA Perspective on International Clinical Trials," Food and Drug Administration, accessed July 26, 2019, https://www.fda.gov/downloads/drugs/newsevents/ucm441250.pdf.

112. Ibid.

113. Ibid.

114. Daniel Castro, "The Rise of Data Poverty in America" (Center for Data Innovation, September 2014), http://www2.datainnovation.org/2014-data-poverty.pdf.

115. "Study Explores Representation of Women in Clinical Trials," American College of Cardiology, April 30, 2018, https://www.acc.org/latest-in-cardiology/articles/2018/04/30/16/43/study-explores-representation-of-women-in-clinical-trials.

116. "Clinical Trials Shed Light on Minority Health," Food and Drug Administration, http://www.fda.gov/ForConsumers/ConsumerUpdates/ucm349063.htm..

117. "Successful Strategies for Engaging Women and Minorities in Clinical Trials," Society for Women's Health Research and U.S. Food and Drug Administration Office of Women's Health, September 2011, http://www.fda.gov/downloads/ScienceResearch/SpecialTopics/WomensHealthResearch/UCM334959.pdf.

118. Sarah Elizabeth Richards, "The DNA Data We Have Is Too White. Scientists Want to Fix That," *Smithsonian*, April 30, 2018, https://www.smithsonianmag.com/science-nature/gene-bank-too-white-180968884/.

119. "Diversity in Clinical Trial Participation," Food and Drug Administration, accessed July 26, 2019, https://www.fda.gov/forpatients/clinicaltrials/ucm407817.htm.

120. "Drug Trials Snapshots," Food and Drug Administration, accessed July 26, 2019, https://www.fda.gov/Drugs/InformationOnDrugs/ucm412998.htm.

121. Chelsea Weidman Burke, "The Importance of Diversity in Clinical Trials (Because Right Now, It's Lacking)," *BioSpace*, October 10, 2018, https://www.biospace.com/article/the-importance-of-diversity-in-clinical-trials-because-right-now-it-s-lacking-/.

122. "FDA Encourages More Participation, Diversity in Clinical Trials," Food and Drug Administration, accessed July 26, 2019, https://www.fda.gov/ForConsumers/ConsumerUpdates/ucm535306.htm.

123. David Beier and George Baeder, "China Set to Accelerate Life Science Innovation," *Forbes*, July 6, 2017, https://www.forbes.com/sites/realspin/2017/07/06/china-set-to-accelerate-life-science-innovation/#59322ee5e73b.

124. Ibid.

125. Robert D. Atkinson, "Health Funding: The Critical Role of Investing in NIH to Boost Health and Lower Costs" (ITIF, March 2019), https://itif.org/printpdf/8377.

126. Joshua New, "Why the United States Needs a National Artificial Intelligence Strategy and What It Should Look Like" (Center for Data Innovation, December 2018), http://www2.datainnovation.org/2018-national-ai-strategy.pdf.; Hal Varian, "Artificial Intelligence, Economics, and Industrial Organization," National Bureau of Economic Research, November 2017, https://www.nber.org/chapters/c14017.pdf.

127. Center for Data Innovation, "U.S. Data Innovation Day 2018: The Future of Data-Driven Medicine," YouTube, September 13, 2018, https://www.youtube.com/watch?v=UqBSeqr_7z8.

128. Adams Nager and Robert D. Atkinson, "The Case for Improving U.S. Computer Science Education" (ITIF, May 2016), http://www2.itif.org/2016-computer-science-education.pdf.

129. Ibid.

130. Ibid.

131. Shaun Michel and James Wittle, "Immigrants Working for U.S. Pharmaceuticals" (George Mason University, August 2014), https://www.immigrationresearch.org/system/files/Immigrants_in_the_Pharma_Industry_Institute_for_Immigration_Research_GMU.pdf.

132. Carolyn Y. Johnson, "Big Pharma Depends on immigrants. It Kept Quiet About Trump's Travel Ban," *The Washington Post*, February 1, 2017, https://www.washingtonpost.com/news/wonk/wp/2017/02/01/big-pharma-depends-on-immigrants-it-kept-quiet-about-the-travel-ban/.

133. Shaun Michel and James Wittle, "Immigrants Working for U.S. Pharmaceuticals" (George Mason University, August 2014), https://www.immigrationresearch.org/system/files/Immigrants_in_the_Pharma_Industry_Institute_for_Immigration_Research_GMU.pdf.; Carolyn Y. Johnson, "Big Pharma Depends on immigrants. It Kept Quiet About Trump's Travel Ban," *The Washington Post*, February 1, 2017, https://www.washingtonpost.com/news/wonk/wp/2017/02/01/big-pharma-depends-on-immigrants-it-kept-quiet-about-the-travel-ban/.

134. Keep STEM Talent Act of 2019, S. 1744, 116th Congress, 2019.

ABOUT THE AUTHOR

Joshua New is a senior policy analyst at the Center for Data Innovation. He has a background in government affairs, policy, and communication. New graduated from American University with degrees in C.L.E.G. (communication, legal institutions, economics, and government) and public communication.

ABOUT THE CENTER FOR DATA INNOVATION

The Center for Data Innovation is the leading global think tank studying the intersection of data, technology, and public policy. With staff in Washington, D.C. and Brussels, the center formulates and promotes pragmatic public policies designed to maximize the benefits of data-driven innovation in the public and private sectors. It educates policymakers and the public about the opportunities and challenges associated with data, as well as technology trends such as open data, artificial intelligence, and the Internet of Things. The Center is a nonprofit, nonpartisan research institute proudly affiliated with the Information Technology and Innovation Foundation.

contact: info@datainnovation.org

datainnovation.org