Data Acquisition, Curation, and Use For a Continuously Learning Health System
A Vital Direction for Health and Health Care

Harlan M. Krumholz, Yale University; **Philip E. Bourne,** National Institutes of Health; **Richard E. Kuntz,** Medtronic; **Harold L. Paz,** Aetna; **Sharon F. Terry,** Genetic Alliance; **Joanne Waldstreicher,** Johnson & Johnson

About the Vital Directions for Health and Health Care Series

This publication is part of the National Academy of Medicine's **Vital Directions for Health and Health Care Initiative,** which called on more than 150 leading researchers, scientists, and policy makers from across the United States to assess and provide expert guidance on 19 priority issues for U.S. health policy. The views presented in this publication and others in the series are those of the authors and do not represent formal consensus positions of the NAM, the National Academies of Sciences, Engineering, and Medicine, or the authors' organizations. Learn more: **nam.edu/VitalDirections.**

Introduction

Increased sharing of health data among all stakeholders in the health system—from patients and advocates to health professionals and medical researchers—is essential for creating a learning health system. Such a system would leverage health data from a variety of sources to meet the challenges of increasingly complex medical decisions and, in the process, create knowledge more efficiently in the service of producing better patient outcomes and less waste. Government agencies, nongovernment organizations (including charitable foundations and disease advocacy organizations), and the research community have taken important strides in recent years toward greater openness of research data and personal health data. In particular, there is increasing movement toward clarifying people's rights to their own health data, promoting standards to ease their access, and providing tools that enable them to exercise their rights. Major challenges remain, however, in overcoming the resistance to data sharing that prevents scientists from learning about clinical trials whose results are unpublished and prevents other people from acquiring and sharing their own health-related data. Those challenges create a need for incentives (financial and otherwise) to create an open-data culture, for changes in laws and regulations to make data sharing easier, for improvement in the infrastructure used for data-sharing, and for investment in research to increase data sharing abilities. Policies promoting a more open system should be evaluated to quantify the transition to a data-sharing ecosystem and the opportunities to improve its effectiveness in promoting clinical quality, patient choice, and scientific progress. Given the scale of the challenges and the potential rewards, a strategic federal initiative that aligns current and future efforts would be one way to

accelerate movement toward a more open, people-centric health system with data sharing at its core.

Topic Overview, Issues, and Trends

Health-related and health-research data are vital resources for clinical care, informed clinical choice, quality improvement, drug and device safety, effectiveness assessment, and scientific discovery. *Health-related data* refers to the four major determinants of health: personal, social, economic, and environmental (ODPHP, no date). Such data are the reagents with which we can produce information to support personal choices about health care, system choices about optimizing medical and public health strategies, and policy choices about laws and regulations. They are the ingredients necessary for medical breakthroughs.

There are formidable impediments—cultural and social as well as technical—to leveraging existing data for the benefit of individuals and society. Because of the incentive structure for data sharing, a prominent impediment is the difficulty in motivating data holders to enable the coalescing and harmonizing of health-related data that reside in disparate venues and formats in the health care and research ecosystems (Murugiah et al., 2016). The ability to access the data is not sufficient to produce benefit; technical advances in analytics and application are also required. Nevertheless, the lack of a way to acquire data easily, securely, and in a useful format is a critical obstacle to producing innovations and improvements in health and health care.

The Institute of Medicine (IOM) (now the National Academy of Medicine) introduced a concept of a learning health system to support transformational change in the fundamental aspects of health and health care (IOM, 2012a). In describing the paradigm shift to a system in which data sharing is the norm rather than the exception, the Office of the National Coordinator for Health Information Technology (ONC), under the aegis of the Department of Health and Human Services (DHHS), defines a learning health system as an ecosystem in which all stakeholders can contribute, share, and analyze data and in which continuous learning cycles encourage the creation of knowledge that can be used by a variety of health information systems (ONC, 2015a). A learning health system has the potential to address some of the most pressing challenges of our current system, including the increasing complexity of medical decisions, the inadequacy and sluggish pace

of acquiring evidence for guiding care, the systemic waste throughout health care delivery, and health disparities and quality shortcomings despite high spending. A learning health system is also intended to expand capacity for knowledge generation, use health information technology (HIT) to propel improvement, configure systems for continuous improvement, and engage patients in working toward better outcomes.

Health-related and research-related data are the substrates for both a learning health system and a vibrant research ecosystem. Such systems require rich, detailed health-related data that are primed to be transformed into useful information at the personal and systems levels. The data must be used optimally in the learning health system for the system to generate useful knowledge for researchers and in turn to leverage this knowledge more quickly and effectively in clinical practice. However, a learning health system remains more an aspiration than a consistent achievement, in part because of an inability to leverage relevant data fully.

Our purpose is to identify the principal opportunities to promote sharing, curation, and use of data for a learning health system and the research ecosystem. In particular, we focus on options for a strategic federal initiative, with additional consideration of the role of others. We articulate the aspirations for data sharing initiatives and metrics for tracking. Three overarching vital directions are needed to create a health and research system that is based on data sharing: change the culture and incentive structures of the health system, encourage people's access to their data by leveraging their established rights to their data, and provide seamless means to curate and produce usable data from disparate sources.

Progress

In recent years, policy makers, organizations, and individuals have advanced efforts to promote the culture and infrastructure needed to support the secure accessibility of health and health care data (Ross and Krumholz, 2013). For example, the companies that are part of the Pharmaceutical Research and Manufacturers of America (PhRMA) have committed to sharing their trial data with researchers (PhRMA, 2013).

There is parallel progress in health care. The spread of digital health data has created the opportunity for people to view, download, and transmit their health

care data and has introduced the possibility of coalescing data from disparate sources. The adoption of electronic health records (EHRs) was an objective of the Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009 and the Federal Health IT Strategic Plan (Henry et al., 2016; ONC, 2014). In 2011, only 28% of hospitals had a basic EHR. By 2015, almost all hospitals (96%) had certified EHR record technology.

Many regions of the country have taken substantial steps to promote data sharing and begin the transition to a learning health system. Regional health information exchanges, despite their limitations, represent progress. An example is the MyHealth Access Network, a nonprofit HIT utility in Tulsa, Oklahoma, supported by ONC as part of the Beacon Communities Program (MyHealth Access Network, 2016). MyHealth supports health-data collection by creating a regional health information exchange that as of 2012 contained the medical records of 1.8 million patients (Tulsa Beacon Community, 2012). The system ensures that every health practitioner who sees a patient has access to the patient's full medical history, and it enables doctors seeing the same patient to coordinate care (Kendrick, 2011).

The promulgation of standards, the implementation of appropriate legislation and regulations, the public attention to what ONC termed information blocking, the growth of public activism regarding health information, and technologic advancements have sped changes in expectations and capabilities (NIHOER, 2016; ONC, 2015b). Information blocking was stated in a congressional report by ONC to occur "when persons or entities knowingly and unreasonably interfere with the exchange or use of electronic health information" (ONC, 2015b). Nevertheless, the focus on common data models, interoperability, and application program interfaces (APIs) and authorization protocols are transforming what is possible with regard to secure health-data movement. The common data models are standards to enable different databases to align elements. APIs—which are software programs, protocols, and tools—are making it easier to move information from one location to another. New standards with an API, such as the Fast Healthcare Interoperability Resources (FHIR), hold the promise of accelerating interoperability. Authorization protocols, such as OAuth 2.0, are providing easier and more secure ways to ensure that appropriate people can gain access to data.

The health care and research worlds are also converging with respect to data flow. An example is the Precision Medicine Initiative's introduction of the Sync-for-Science concept. That effort seeks to engage people in acquiring their health-related data, including data from EHRs, and transmitting the data into research databases (PMIWG, 2015).

National legislation and guidance from ONC and DHHS are accelerating the transformational change to a digital health-data environment (ONC, 2015a). The 1996 Health Insurance Portability and Accountability Act (HIPAA) made clear that Americans have a right to access their health data, to have an accounting of their health information, and to correct or amend their health information (HealthIT.gov, no date a). The HITECH Act, a part of the 2009 American Recovery and Reinvestment Act, made clear that Americans have a right to acquire their personal health information (PHI) in an electronic format; as a result, gatekeepers to those data are obliged to provide the data on request (DHHS OCR, no date). The legislation stated that a person can be charged only the labor cost. The DHHS Office for Civil Rights (OCR) guidance states that, "while a covered entity is not required to purchase new software or equipment in order to accommodate every possible individual request, the covered entity must have the capability to provide some form of electronic copy of PHI maintained electronically" (HIPD, no date). Progress with regard to fees was also made with new guidance from OCR released in early 2016. The guidance now states that "a covered entity may charge individuals a flat fee for all standard requests for electronic copies of PHI maintained electronically, provided the fee does not exceed $6.50, inclusive of all labor, supplies, and any applicable postage" (HIPD, no date).

ONC released a Shared Nationwide Interoperability Roadmap in 2015 (ONC, 2015a). The short-term goals (for 2015–2017) focus on "sending, receiving, finding and using priority data domains to improve health care quality and outcomes." The longer-term goals (for 2018–2020) address the need "to expand data sources and users." The even longer-term goals (for 2021–2024) seek broadly to "achieve nationwide interoperability to enable a learning health care system, with the person at the center of a system that can continuously improve care, public health, and science through real-time data access." ONC also released a federal HIT

strategic plan for 2015–2020, which stated that the mission is to "improve the health and well-being of individuals and communities through the use of technology and health information that is accessible when and where it matters most" (ONC, 2014).

Many federal agencies are sharing data at an increasing pace. For example, the Centers for Medicare & Medicaid Services (CMS) began releasing data several years ago and has progressed quickly to sharing information of many kinds, including data on hospital discharges, physician volumes, drug prescribing, and durable medical equipment (CMS, no date; Ornstein, 2016). Moreover, CMS is building APIs that will enable Medicare beneficiaries to connect their CMS data to personal applications in ever easier and more expeditious fashion.

The expansion of alternative payment models (APMs) makes health data sharing more important and creates new incentives to do so. The APMs are likely to grow more rapidly with the advent of the Medicare Access and CHIP Reauthorization Act of 2015, which introduced a Quality Payment Program. APMs serve as an impetus for data sharing, as the move away from a fee-for-service (FFS) model creates a need for longitudinal patient data to enable effective and efficient care over a patient's lifetime. In a FFS model, institutions could get by with data about individual episodes of care; in APMs, institutions increasingly need HIT systems that integrate data over time and enable sharing with other institutions as needed to provide longitudinal care and act to promote health. For example, Blue Cross Blue Shield of Massachusetts launched an APM in 2009 called the Alternative Quality Contract, which pays a fixed amount, linked to quality measures, for each patient during a specific period. To manage population health with multiple providers in such a system, Blue Cross created a data-reporting system that helps physicians with medical management and provides a mechanism to share best practices and monitor quality measures. The infrastructure in the system could serve as the base for a broader data-sharing system.

Progress is being promoted by many nongovernment organizations. DirectTrust is a nonprofit collaborative that consists of providers that seek methods for a secure, interoperable health information exchange via the Direct message protocols (DirectTrust, 2012). The Argonaut Project is a collaborative effort to facilitate data sharing by using FHIR (FHIR, 2015). The

CommonWell Health Alliance is organizing HIT companies and other stakeholders to promote interoperability (CommonWell Health Alliance, no date). Moreover, companies that provide 90% of the country's EHRs and several large health systems have signed the ONC Interoperability Pledge and committed to consumer access, no blocking, ensuring transparency, and implementing standards (HealthIT.gov, no date b).

On the research side, there have been advances in the commitment of influential organizations to mandate data sharing in research. IOM convened meetings over the last several years to discuss data sharing in science and made strong recommendations for promoting progress toward a culture of open science. Many data holders, including PhRMA, are committed to sharing their data, and consortia, individual academic groups, companies, and others have established mechanisms to vet proposals and provide access to their clinical-trial assets (PhRMA, 2013).

Funders are increasingly linking financial support with data sharing. Organizations that include the National Institutes of Health (NIH) and the Patient-Centered Outcomes Research Institute have mandated some forms of data sharing as a condition of funding (Goodman and Krumholz, 2015). They have developed platforms for sharing, are investing in the concept of a data commons, and are committed to testing policy and infrastructure approaches. The Wellcome Trust is seeking to identify structures to enable sharing, stating as its aim "to ensure that the data generated by the research we support is managed and shared in a way that maximizes the benefit to the public" (Wellcome Trust, no date a). Wellcome is also launching a new publishing platform, which will encourage publication and data sharing (Wellcome Trust, no date b). Leaders of advocacy organizations have formally convened to propose shared principles that are based on the recommendations.

It is of particular note that in 2014, the Bill & Melinda Gates Foundation promulgated one of the strongest requirements for sharing, making it a contingency of being funded (Straumsheim, 2014). The foundation states that "information generated during the course of our investment activities—in the form of research studies, data sets, evaluation results, investment results and strategy-related analytics—is significant public good. Access to this information is important for accountability, provides valuable learning to the sectors

that we support, will facilitate faster and more well-informed decision making, and contributes to achieving the impact we seek" (Bill & Melinda Gates Foundation, no date a). The foundation also adopted an open-access policy that "enables the unrestricted access and reuse of all . . . peer-reviewed published research funded . . . by the foundation, including any underlying data sets" (Bill & Melinda Gates Foundation, no date b).

The International Committee of Medical Journal Editors, on January 20, 2016, released a proposal that could change the landscape of research data sharing (Taichman et al., 2016). The committee stated the belief that there is "an ethical obligation to responsibly share data generated by interventional clinical trials." It proposed requiring authors "to share with others the deidentified individual-patient data (IPD) underlying the results presented in the article (including tables, figures, and appendices or supplementary material) no later than 6 months after publication. The data underlying the results are defined as the IPD required to reproduce the article's findings, including necessary metadata." The committee received more than 300 comments and is considering whether to adopt the policy or modify it.

Challenges

Despite that progress, data sharing is not easy or normative in health care or clinical research. There are daunting obstacles to individuals in accessing their own health care data, let alone data in a useful form. Sharing among researchers, not to mention broader access, is still relatively uncommon, although a recent study provides evidence of its benefit (McKiernan et al., 2016).

Clinicians are often missing clinical information on their patients, and longitudinal information on patients is difficult and expensive to obtain (Smith et al., 2005). Health care systems that seek to improve are stymied by the lack of longitudinal data, which limits them to a partial view of patients. In addition, information on the safety and effectiveness of some approved drugs and devices is incomplete, and this may undermine surveillance efforts (Brookings Institution, 2015).

Scientists are often blocked from accessing research data generated by others even when the work was funded by federal agencies. The IOM report *Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risks*

states the problem succinctly: "Vast amounts of data are generated over the course of a clinical trial; however, a large portion of these data is never published in peer-reviewed journals" (IOM, 2015a). The consequence of this scientific culture is inefficiency and irreproducibility. The incomplete, inadequate, and even absent harvest of research data, even those generated with public funds, wastes research investment and dishonors the contributions of research participants. Moreover, it slows scientific progress and impedes the self-correcting nature of good science (Silberzahn and Uhlmann, 2015). Academic institutions and their organizations have been relatively quiet about data sharing. For example, the authors of 88% of NIH-funded journal articles did not deposit their datasets into known repositories, and this keeps the data "invisible" (Read et al., 2015).

Despite federal regulations, the path to data access is often not easy. Many institutions do not provide seamless ways to transmit or download data. Despite the advocacy of the OpenNotes movement to make clinical notes visible to patients, many institutions do not share this digital information without substantial effort by patients. Some individuals and organizations have formed coalitions to bring attention to the issue, such as Free the Data, Get My Health Data, and Get My Data [1, 2, 3]. The coalitions are making slow headway, and there are reports of resistance by those who are concerned that HIPAA prevents people from accessing their health information (which is false) or who are not clear about the various secure transmission mechanisms, such as Direct (DirectTrust, 2012; Evans, 2016; Lohr, 2011). In addition, participants and potential participants in clinical trials are often unable to facilitate sharing of clinical data. Many people do not understand the power of sharing their own health data and are therefore not creating the demand for their data. It is noteworthy that Pfizer now shares data collected in clinical trials with patient participants, both providing patients with nontechnical summaries of trial findings and using Blue Button technology to allow patients to access all collected medical data directly and integrate them into EHRs (Pfizer, 2016).

For any data sharing to be useful, it will first be necessary to ensure that health-data records are trustworthy enough and interoperable among different systems. Improving the quality of notes is also relevant to written records, although some issues are specific to

EHRs. There are reports of egregious errors and growing verbiage in electronic medical records, especially as health providers resort to copy-and-paste to fill out the records (Hirschtick, 2006). A 2012 IOM report, *Health IT and Patient Safety: Building Safer Systems for Better Care*, found that poor implementation and use of HIT could lead to new hazards, such as dosing errors or delays in the detection of illnesses (IOM, 2012b). A 2013 report published by members of the American College of Emergency Physicians identified the need for EHR users to have a systematic process to provide comments about potential safety problems and other issues with the EHR systems—a departure from the current system wherein some EHR vendors prohibit users from sharing potential dangers, even in academic publications (Farley et al., 2013). Despite the challenges, there remains much that is trustworthy and reliable in EHRs.

The biggest issue is that progress is not fast enough. For data holders, sharing can represent the loss of a valued asset and the exposure of their work to the scrutiny of others, and the incentives of data holders are not always fully aligned with those of patients and other researchers and physicians. Part of the problem stems from the cost structure, wherein data sharing requires both upfront and continuing spending on infrastructure, administration, standardization, and human resources (Wilhelm et al., 2014). And of course, data holders face substantial opportunity costs—the time and resources spent on sharing data that would otherwise have gone to conducting new research, running analyses, and generating new data. One particular data-sharing project for Alzheimer's disease research found that 10–15% of total costs and 15% of investigators' time was spent on data-sharing activities (Wilhelm et al., 2014). Given that more comprehensive data-sharing projects will impose commensurately higher costs on the data holder and that the benefits will be spread among all parties, some researchers find themselves supporting data sharing for others without sharing their own data.

Many institutional data holders face a public-goods problem with data sharing. Individual data holders will not capture the full social benefits of their own data sharing and will thus underinvest in sharing even as all parties benefit when a single data holder decides to share (Hall, 2014). In the language of economics, data sharing has positive externalities but internalized costs, and this leads to an undersupply of shared data. Mark Hall illustrates that reality with a small-scale

example of a patient who has seen four doctors and is heading to a fifth; only the fifth doctor and the patient benefit from the first four doctors' data sharing (Hall, 2014). It cannot be assumed that the five doctors share patients in the same proportion, and the doctors will not necessarily agree to a reciprocal, quid pro quo data-sharing agreement, inasmuch as different doctors have different incentives to share data. Data sharing in connection with clinical trials presents a similar conundrum. A solution to the problem will require a realignment of incentives that enables doctors and researchers to focus on the best outcomes for patients without having to bear a disproportionate share of the costs.

Even those who seek to share data often encounter problems. For example, the IOM committee identified infrastructure, technology, workforce, and sustainability as key challenges in clinical-trial data sharing—issues that apply to all types of health care data sharing (IOM, 2015a). However, the IOM committee that studied the issue could not find a case of "harm" to data holders in data sharing.

In health systems, the sharing of data can enhance options for patients and reduce barriers to changing providers. The issues of access and security are ever-present concerns. The need to respect privacy concerns associated with a person's health-related data and the need to obtain permission, as appropriate, are equally important. The challenge of inadequate metadata, including documentation, impedes progress. Combining datasets that do not have common data models or that have inconsistently applied common models—and duplicative, sometimes conflicting, information—creates problems in use. The timely updating of data that continue to accumulate and the correction of errors remain problematic. High-quality, longitudinal health-related data remain missing, particularly data generated from devices and responses to patient-reported measures and surveys.

Another issue is the movement of health care data without patients' permission. The Shared Nationwide Interoperability Roadmap states that the goal is a system with the patient at the center (ONC, 2015a). However, massive amounts of data are moving without people at the center. One company claims to have some 300 million EHRs—but without the people's permission (Lohr, 2016). Many companies traffic in a health-data economy, but patients are rarely asked to provide permission for movement of their records. Permission is not always possible, and there are permitted uses

and disclosures, but it is possible that there can be greater focus on making it easy for people to be involved in decisions about their data.

The issue of permission is also bound to the issue of combining datasets. A 2012 paper in *Nature Reviews Genetics* identified the need to merge EHR data among regions to maximize the gains for research. The authors argued that true data interoperability would require "the development and implementation of standards and clinical-content models for the unambiguous representation and exchange of clinical meaning" (Jensen et al., 2012). All data-sharing activities today proceed with the institution at the center. As long as Institution A shares data with Institution B without involving the person to whom the data belong, there will be duplicative and incomplete data and difficulty in collecting them longitudinally. However, systems that are centered on the person allow much clearer and cleaner data sharing, much as financial systems allow people to move funds among financial accounts, instruments, and institutions. The person gives permission and manages issues surrounding identity. Such systems in health information management would produce the same benefits.

The size and complexity of the data require new techniques if the data are to yield important insights. Emerging big-data tools, which have proved valuable in other fields, have little utility without useful data. In the research arena, progress is slow; many studies are never published or reported—at least within a reasonable timeframe—and data sharing is an infrequent and often unavailable option (Ross et al., 2012). The computational burden may also be large and require new investment. Data sharing involves considerable costs, such as the costs of developing an infrastructure, curating the data, supporting security measures, and making operations transparent for clinical research sharing. Who would pay for such systems and how the return on investment would be measured are still unclear. Perhaps the most critical issues to be addressed are how the systems can be sustainable and who should bear the burden of the costs.

Priority Considerations

The following considerations apply to the sharing of research data and health-related data (most often with patient permission). The overall goal is to increase the capacity of the health care and medical-research

enterprises to enable efficient, secure, and permission-based sharing of data—and for people to be involved, to the extent possible, in decisions about their data. Moreover, in cases in which detailed consent is not possible, there is an imperative to remain attentive to privacy concerns. The considerations are in five main categories: foster a culture of data sharing, improve incentives for data sharing, create legal and regulatory tailwinds for data sharing, strengthen the infrastructure for data sharing, and invest in research and training related to data sharing.

Foster a Culture of Data Sharing

Improvements in data sharing in health care and science start with fostering a culture. For data sharing and its use to spread, the culture of health care and science will need to evolve in such a way that refusal or inability to share is understood as against the best interests of individuals and society. In health care, there should be a broad understanding of the rights of a person to view, download and access, and transmit or share his or her own health data, although it is important to remember that people retain the right not to share data. In research, there should be an understanding that good science and good scientific citizenship require that participant-level data be available for evaluation and reuse. Cooperative efforts among government, academic institutions, industry, consumer-advocacy organizations, and experts in science, health care, and ethics could set common expectations and build on foundational consensus documents, such as those produced by IOM. Statements by DHHS Secretary Sylvia Burwell and NIH Director Francis Collins have demonstrated strong support for data sharing (Bowman, 2016; Healy, 2014). Such leadership and expectations need to be internalized throughout the health care and scientific communities.

There is a need to attend to the culture in medicine that has typically marginalized the right of people to be able to access their health records, failed to emphasize the potential for data to create smarter and more responsive health care delivery, and created the notion that investigators have discretion over sharing research results and data. An initiative directed toward fostering a culture of data sharing is warranted. The following proposals would help to kick-start the shift to a culture of data sharing:

- Engage social scientists to define cultural and economic forces that support the status quo.
- Define benefits of data sharing for different stakeholders.
- Identify levers that will change cultural norms regarding data sharing, recognizing that much of that change will come from new incentive models.
- Support working groups to develop clear articulation of the societal value of data sharing.
- Educate the public about data sharing, being attentive to privacy issues, including cases that illustrate the value.
- Define interventions to change the culture regarding data sharing in health care and medical research.

Improve Incentives for Data Sharing

Behaviors that are counter to a culture of data sharing are reinforced by current incentives. Those incentives benefit those who sequester data assets, uphold barriers that prevent people from accessing their records, deny organizations the ability to leverage data, and prevent scientists from sharing data. The evolution to a culture of data sharing will require a shift in the incentives:

- Develop rewards for data sharing and develop penalties for not sharing data.
- Require, to the greatest extent possible, the sharing of trial data with the publication of trial results.
- Encourage publishers to require that data be deposited at the time of publication.
- Provide reimbursement benefit for health systems that facilitate sharing with patients and researchers.
- Provide incentives to companies that have data-sharing programs.
- Give credit for data sharing and downstream use in the process for academic promotion.
- Seek solutions through challenges, such as the DHHS Move Health Data Forward Challenge.
- Publicly report metrics on ease of data accessibility for patients at the hospital, health-system, or office level.

Create Legal and Regulatory Tailwinds for Data Sharing

Legal and regulatory actions by the government will be important levers for change. Interest in data sharing is relevant to many federal agencies and departments, including ONC, CMS, the Food and Drug Administration (FDA), NIH, the Health Resources and Services Administration, the Agency for Healthcare Research and Quality, the Department of Defense, the Department of Veterans Affairs, and the Centers for Disease Control and Prevention. The IOM report *Vital Signs: Core Metrics for Health and Health Care Progress* issued a clarion call for coordination and alignment among multiple government agencies in the context of identifying core metrics for measuring health and health care progress (IOM, 2015b). The report argues that opportunities are lost when data collected in one program do not work synergistically with data in another program and when data are not used to create new knowledge. Drawing on the example of the IOM *Vital Signs* report, the alignment of many federal agencies and departments in support of data sharing is critical for providing momentum to change the culture and behaviors in the research environment. In fact, as exemplified in the federal HIT strategic plan, there is already collaboration among federal organizations.

- Establish discussion, including consumers, on permitted uses and disclosures related to which data can be shared without people's explicit permission and provide guidance on informed-consent language.
- Continue to link requirements to facilitate sharing with funding, certification, and approval.
- Continue to promote and harmonize federal standards relevant to data sharing.
- Continue to extend federal standards for ownership, security, and privacy of health care data.
- Continually evaluate regulations, such as those based on HIPAA, following the guidelines of the IOM report *Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health Through Research* (IOM, 2009).
- Require a unique medical-device identifier in every relevant electronic medical record and on administrative claims, building on CMS and FDA recommendations (Rubenfire, 2016).
- Encourage use of standardized authentication systems for patient portal access, using the OAuth 2.0 authorization standard as a model.
- Investigate the value of new approaches, such as FHIR, and promote successful models, highlighting not only the approach but best practices in implementation.

- Promote the provision of information to people about their data rights.
- Develop mechanisms for easy public reporting of instances of information blocking.
- Penalize information-blocking.
- Establish an honor roll for health-related companies that have exemplary sharing policies.
- Penalize academic institutions that do not share data produced with federally funded grants.
- Highlight publicly the data-sharing performance of academic institutions.
- Provide benefits for data sharing in the drug-approval and device-approval process.
- Require data sharing (following the IOM recommendations) for studies that use public funds.
- Support the idea of data sharing related to trials published in journals.

Strengthen the Infrastructure for Data Sharing

As noted in the IOM report, platforms for storing and managing trial data efficiently are inadequate. The lack of infrastructure applies equally to a variety of data assets in health care and science, including personal health information and basic-research data.

- Convene stakeholders and seek common requirements for infrastructure.
- Investigate economies of scale and benefits of competition.
- Define particular needs of different stakeholder groups.
- Identify opportunities for joint ventures between aligned groups, including federal agencies and departments.
- Investigate sustainable business models for data-sharing infrastructure.
- Investigate government solutions for data-sharing infrastructure.
- Define minimal costs of high-quality data sharing in different venues.
- Develop means of promoting FAIR (find, access, interoperate, reuse) principles (Wilkinson et al., 2016).
- Create standards that guarantee people access to their own research data.
- Create standards for informed consent that consider reuse of research data.
- Invest in the human capital necessary to advance an ecosystem that promotes data sharing.

- Continue to open federal databases to the public through APIs such as FHIR (or other suitable means).
- Continue development and dissemination of ontologies (the classes, properties, and relationships between class members with which to model health data sharing).
- Investigate a unique national patient identifier and other strategies to combine a person's health-related data.
- Support the development and implementation of participant-centric data-sharing solutions.

Increase Capability by Investing in Research on Data Sharing

Success in optimizing the organization and use of data to achieve better health and health care will depend on the capability of generating knowledge. The capability to do so will require investment in research that is germane to data sharing. We need to apply what we know while developing more fully the science that underlies successful and sustainable data sharing in health care and science.

The issue of data sharing has technological, computational, organizational, economic, and social dimensions, all of which require study. Research investment should span data science, implementation science, management science, network science, economics, law, and health policy.

Also important is the scope of research in data science. Designing a new assay is considered scientific, but developing a new genomic alignment algorithm or approach for data interoperability is not. To embrace data-driven health care, we need a culture shift in what is considered science, as distinct from infrastructure, from a computational perspective.

- Develop novel approaches to deidentification and privacy concerns.
- Support national surveys of the public's views on data sharing in health care and science.
- Support funding for primary informatics research that is relevant to data sharing.
- Develop analytics suited to shared data and their particular challenges.
- Develop methods that address data access and security.
- Develop methods to enhance data sharing for people who have limited technical ability, health literacy, or access to technology.

- Develop platforms that increase the efficiency and transparency of sharing.
- Develop tools and methods to support infrastructure.
- Test, strengthen, and refine or improve common data models.
- Develop new models of academic credit for sharing data.
- Develop analytics tuned to issues peculiar to data sharing.
- Develop strategies that lower the cost of data sharing.
- Test strategies for enforcing data-sharing policies.
- Investigate benefits, risks, and costs associated with data-sharing, especially as behavior evolves.
- Investigate ethical underpinnings of the imperative to share data for societal benefit.
- Investigate state-based initiatives to assess effects of data sharing, and use states as laboratories.
- Build on evidence-based methods in other fields; pilot-test strategies for engaging the public.
- Evaluate the quality of data being shared and standards for sharing.
- Provide funding mechanisms for data sharing.
- Value those who contribute to data science as we do other researchers and health care professionals.

Options for Strategic Federal Initiatives

Strategic federal initiatives are needed for issues whose substantial consequences span multiple levels of influence. An overarching strategy to promote sharing, curation, and use of data to improve health and health care must address key impediments to progress and promote a view of a better future while articulating the features of that future. The recommendations above focus attention on linchpins in the movement toward data sharing: culture, incentives, infrastructure, and capability. Only the federal government, with its many agencies and departments, can provide the impetus for each of those to enlist the support of other key stakeholders nationwide. Such a pathway would build on successful initiatives that are making data sharing better, faster, and less expensive—strengthening them and enabling data sharing and transparency to be vital parts of efforts to improve health care and science in tandem, invigorating a data economy, and producing marked societal gains. Many of the efforts are already under way in the federal government, and

it is important to avoid duplication. Such an initiative could be undertaken by DHHS with the US Chief Technology Officer and would be best accomplished as a White House initiative spanning the government. It would also seek to support market forces in leveraging government efforts by creating products that facilitate the use of increasingly available data. The government has the power to recognize achievements, promote education about rights and laws, institute standards, penalize infractions, and protect individuals. This topic is thus primed for a strategic federal initiative, building on and strengthening existing efforts, to accelerate progress toward an era in which digital health-related data could fulfill their role in creating smarter, more personalized health care and more rapid, timely, and efficient science. DHHS should conduct participant-centric, citizen science–based pilots based on digital health data to accelerate learning and begin real-world implementation.

Potential Metrics

Increasing access to health-related data, with people at the center, and producing tools to leverage the data as part of a learning health system could have dramatic effects. The more people own their own health and wellness data, the more likely it is that they will be able to act on them to create better value for themselves. It should be possible to leverage digital data fully to ensure that individual health care decisions are informed by all the data; that, with permission, the data could be used for research and system improvement; and that the data could increase transparency in health care and be an impetus toward improved quality and reduced waste. The potential knowledge trapped within those digital data should be released to propel health care toward more effective and efficient practice in such a way that we could save the time and resources currently devoted to chasing data sources and repeating clinical testing. Medicine would improve if clinicians knew that patients would see their work and could easily share it with other experts for second opinions. Greater data availability could enable people to see how thousands of others who have similar clinical characteristics and backgrounds responded to different treatment paths and then have an evidence-based discussion with their doctors before embarking on a specific treatment plan. It is possible that if people had

a say in how their data were used and were positioned to enable higher-quality, more timely, and more comprehensive data to fuel new insights, it could help other people who had similar problems. Health systems and other health care providers could use the data to redesign care and improve results. Scientists could perceive their data as a public good and would share generously, seeking to accelerate progress and finding ways to reward most those who enable others to produce important insights. Savings could be achieved if we sought full harvesting of data generated through research and provided opportunities for reexamination, reanalysis, and reinterpretation of study data to promote public discussion in search of truth. The quality of science could increase if researchers knew that others would view their work, their operating manuals, and their processes.

Interventions that aspire to promote data sharing as a means of improving health care should be evaluated by measures that assess progress toward the goal and monitor for unintended adverse consequences. Leading indicators can signal whether other forces are promoting or impeding progress and results. The metrics should be used to assess progress in enabling people to obtain and use their health data, enabling organizations to share and use their data, and enabling researchers to report and share their data. The development of metrics requires input from stakeholders, data sources to enable the calculations, and specifications that promote a reflection of the domain under assessment. Details aside, we present below a sampling of metrics that could be used to track progress in data sharing:

- Percentage of late stage clinical trials by funder with complete and accurate reporting in clinicaltrials.gov within 12 months and publication within 18 months of completion.
- Percentage of clinical trials by academic center reported within 12 months and published within 18 months of completion.
- Percentage of nation's hospitals that have Blue Button capability, the ability of patients to view and download their personal health records.
- Percentage of 1,000 largest physician offices that make it possible for patients to view, download, and transmit their EHR information.

- Percentage of nation's 100 largest hospitals to move data by FHIR API with a common data standard.
- Percentage of patients in nation's hospitals who have patient portals.
- Percentage of hospitals and offices that have high-quality data from patient portals, according to high-quality data standards.
- Percentage of academic institutions that commit to incorporate data sharing into decisions on individual promotions.
- Percentage of academic institutions that have data-sharing initiatives.
- Percentage of federally funded medical-research grantees who report results in a public venue within 12 months of finishing their studies.
- Number of publications per year that are based on NIH-shared datasets.
- Number of publications from prominent data sharing efforts.
- Number of complaints about information blocking and its root causes.
- Number of initiatives for data sharing throughout federal agencies.

Conclusion

Data sharing, data curation, and data use for a continuously learning health system hold great potential for promoting better engagement by people in their health and health care, better care, less waste, better outcomes, and greater progress toward medical breakthroughs. To move forward, there are three vital directions. The first is a change in the culture and incentive structure of the health system and research enterprise to move away from a status quo anchored in an environment that offers little opportunity for data sharing. The inefficiencies, errors, restrictions, duplication, and waste imposed by barriers to sharing and use of digital health-related data cost lives and resources. The second direction is to encourage people's access to their data by clarifying and strengthening their rights to their data. This would require changes in regulatory structures and the creation of the tools and infrastructure needed for patients to put their data to work for them. Building on the first two, the third and final direction is to provide seamless means to curate and produce usable data from disparate sources to promote opportunities for improvements in health and

health care. Data can fuel the learning health system of the future; but as long as data remain in discrete silos, people will be unable to leverage their own data fully to create maximum value for their own health. Moving toward an enlightened system that grows smarter with the accumulation of data will require unprecedented levels of collaboration among and communication between all stakeholders in the health system. Such a grand strategy for change offers an ideal opportunity for government facilitation and support because these changes are likely to yield an immense return on investment for society.

Summary Recommendations for Vital Directions

1. Foster a culture of data sharing.
2. Create the operational functionality for data sharing.
3. Build the continuous data-sharing improvement capacity.

Notes

1. Free The Data. http://www.free-the-data.org/.
2. Get My Data. http://getmydata.org/#about.
3. Get My Health Data. https://getmyhealthdata.org/.

References

Bill & Melinda Gates Foundation. No date a. Bill & Melinda Gates Foundation open access policy. Available at http://www.gatesfoundation.org/How-We-Work/General-Information/Open-Ac¬cess-Policy (accessed August 25, 2016).

Bill & Melinda Gates Foundation. No date b. Information sharing appoach. Available at http://www.gatesfoundation.org/how-we-work/general-information/information-sharing-approach (ac¬cessed August 25, 2016).

Bowman, D. 2016. Sylvia Mathews Burwell: Work remains to make healthcare system open. Available at http://www.fiercehealthit.com/story/sylvia-mathews-burwell-work-remains-make-healthcare-system-open/2016-05-10 (accessed August 25, 2016).

The Brookings Institution. 2015. *Strenthening patient care: Building an effective national medical device surveillance system.* Available at http://www.fda.gov/downloads/aboutfda/centersoffices/officeof-medicalprod-uctsandtobacco/cdrh/cdrhreports/ucm435112.pdf (accessed August 25, 2016).

CMS (Centers for Medicare & Medicaid Services). No date. CMS data navigator. Available at https://dnav.cms.gov/ (accessed August 25, 2016).

CommonWell Health Alliance. No date. Why CommonWell Health Alliance. Available at http://www.commonwellalliance.org/ (accessed August 25, 2016).

DHHS OCR (Department of Health and Human Services Office for Civil Rights). No date. HITECH Act enforcement interim final rule. Available at http://www.hhs.gov/hipaa/for-pro¬fessionals/special-topics/HITECH-act-enforcement-interim-final-rule/index.html (accessed August 25, 2016).

DirectTrust. 2012. What is DirectTrust? Available at https://www.directtrust.org/about-directtrust/ (accessed August 25, 2016).

Evans, B. 2016. Barbarians at the gate: Consumer-driven health data commons and the transformation of citizen science. *American Journal of Law and Medicine* 42(4).

Farley, F., K. Baumlin, A. Hamedani, D.S. Cheung, M.R. Edwards, D.C. Fuller, N. Genes, R.T. Griffey, J.J. Kelly, J.C. McClay, J. Nielson, M.P. phelan, J.S. Shapiro, S. Stone-Griffin, and J.M. Pines. 2013. Quality and safety implications of emergency department information systems. *Annals of Emergency Medicine* 62(4):399-407.

FHIR (Fast Health Interoperability Resources). 2015. The Argonaut Project. Available at http://hl7.org/fhir/2015Jan/argonauts.html (accessed August 25, 2016).

Goodman, S., and H. Krumholz. 2015. Open science: PCORI's efforts to make study results and data more widely available. Available at http://www.pcori.org/blog/open-science-pcoris-efforts-make-study-results-and-data-more-widely-available (accessed August 25, 2016).

Hall, M. 2014. Property, Privacy and the Pursuit of In¬tegrated Electronic Medical Records. Wake Forest University Legal Studies Paper 1334963. doi: 10.2139/ssrn.1334963.

HealthIT.gov. No date a. Your health information rights. Available at https://www.healthit.gov/patients-families/your-health-information-rights (accessed August 25, 2016).

HealthIT.gov. No date b. Interoperability pledge. Available at https://www.healthit.gov/commitment (accessed August 25, 2016).

Healy, M. 2014. Big data, meet big money: NIH funds centers to crunch health data. Available at http://www.latimes.com/science/sciencenow/la-sci-sn-big-data-money-20141009-story.html (accessed August 25, 2016).

Henry, J., Y. Pylypchuk, T. Searcy, and V. Patel. 2016. Adoption of Electronic Health Record Systems among U.S. Non-Federal Acute Care Hospitals: 2008- 2015. ONC Data Brief 35, May. Available at http://dashboard.healthit.gov/evaluations/data-briefs/non-federal-acute-care-hospital-ehr-adoption-2008-2015.php (accessed August 25, 2016).

HIPD (Health Information Privacy Division). No date. Individuals' right under HIPAA to access their health information 45 CFR § 164.524. Available at http://www.hhs.gov/hipaa/for-professionals/privacy/guidance/access/ (accessed August 25, 2016).

Hirschtick, R. 2006. Copy-and-paste. *Journal of the American Medical Association* 295(20):2335-2336.

IOM (Institute of Medicine). 2009. *Beyond the HIPAA privacy rule: Enhancing privacy, improving health through research.* Washington, DC: The National Academies Press.

IOM. 2012a. *Report brief: Best care at lower cost: The path to continuously learning health care in America.* Available at http://www.nationalacade¬mies.org/hmd/~/media/Files/Report%20Files/2012/Best-Care/BestCareReportBrief.pdf (accessed Au¬gust 25, 2016).

IOM. 2012b. *Health IT and patient safety: Building safer systems for better care.* Washington, DC: The National Academies Press.

IOM. 2015a. *Sharing clinical trial data: Maximizing benefits, minimizing risk.* Washington, DC: The National Academies Press.

IOM. 2015b. *Vital signs: Core metrics for health and health care progress.* Washington, DC: The National Academies Press.

Jensen, P., L. Jensen, and S. Brunak. 2012. Mining electronic health records: Towards better research applications and clinical care. *Nature Reviews Genetics* 13:395-405.

Kendrick, D. 2011. The Beacon communities at one year: The Tulsa experience. Available at http://healthaffairs.org/blog/2011/06/01/the-beacon-communities-at-one-year-the-tulsa-experience/ (accessed August 25, 2016).

Lohr, S. 2011. U.S. tries open-source model for health data systems. *New York Times,* February 2. Available at http://bits.blogs.nytimes.com/2011/02/02/u-s-tries-open-source-model-for-health-data-systems/ (accessed August 25, 2016).

Lohr, S. 2016. IBM buys medical analytics company for $2.6 billion. *New York Times,* February 19, p. B3.

McKiernan, E. C., P. E. Bourne, C. T. Brown, S. Buck, A. Kenall, J. Lin, D. McDougall, B. A. Nosek, K. Ram, C. K. Soderberg, J. R. Spies, K. Thaney, A. Updegrove, K. H.Woo, and T. Yarkoni. 2016. Point of view: How open science helps researchers succeed. *eLife* 5:e16800.

Murugiah, K., J. D. Ritchie, N. R. Desai, J. S. Ross, and H. M. Krumholz. 2016. Availability of clinical trial data from industry-sponsored cardiovascular trials. *Journal of the American Heart Association* 5(4):e003307.

MyHealth Access Network. 2016. Available at http://myhealthaccess.net/who-we-are/ (accessed August 25, 2016).

NIHOER (National Institutes of Health Office of Extramural Research). 2016. NIH sharing policies and related guidance on NIH-funded research resources. Available at https://grants.nih.gov/policy/sharing.htm (accessed August 25, 2016).

ODPHP (Office of Disease Prevention and Health Promotion). No date. Determinants of health. Available at https://www.healthypeople.gov/2020/about/foundation-health-measures/Determinants-of-Health (accessed August 25, 2016).

ONC (Office of the National Coordinator for Health Information Technology). 2014. Federal Health IT Strategic Plan: 2015-2010. Available at http://dashboard.healthit.gov/strategic-plan/federal-health-it-strategic-plan-2015-2020.php (accessed on August 25, 2016).

ONC. 2015a. Connecting health and care for the nation. A shared nationwide interoperability road¬map. Available at https://www.healthit.gov/sites/default/files/hie-interoperability/nationwide-interoperability-roadmap-final-version-1.0.pdf (accessed August 25, 2016).

ONC. 2015b. Report on health information blocking. Available at https://www.healthit.gov/sites/default/files/reports/info_blocking_040915.pdf (accessed August 25, 2016).

Ornstein, C. 2016. What Feds' push to share health data means for patients. Available at http://www.scpr.org/news/2016/05/09/60446/what-feds-push-to-share-health-data-means-for-pati/ (accessed August 25, 2016).

Pfizer. 2016. Returning clinical data to patients. Available at http:// www.pfizer.com/research/clinical_trials/trial_data_and_results/data_to_patients (accessed August 25, 2016).

PhRMA. 2013. Principles for responsible clinical trial data sharing. Available at http://www.phrma.org/phrmapedia/responsible-clinical-trial-data-sharing (accessed August 25, 2016).

PMIWG (Precision Medicine Initiative Working Group). 2015. The Precision Medicine Initiative Cohort Program—Building a research foundation for 21st cen¬tury medicine. Available at http://acd.od.nih.gov/reports/PMI_WG_report_2015-09-17-Final.pdf (accessed August 25, 2016).

Read, K. B., J. R. Sheehan, M. F. Huerta, L. S. Knecht, J. G. Mork, B. L. Humphreys, and NIH Big Data Annotator Group. 2015. Sizing the problem of improving discovery and access to NIH-funded data: A preliminary study. *PLoS One* 10(7):e0132735.

Ross, J. S., and H. M. Krumholz. 2013. Ushering in a new era of open science through data sharing: The wall must come down. *Journal of the American Medical Association* 309(13):1355-1356.

Ross, J. S., T. Tse, D. A. Zarin, H. Xu, L. Zhou, and H. M. Krumholz. 2012. Publication of NIH funded trials registered in ClinicalTrials.gov: Cross sectional analysis. *British Medical Journal* 344:d7292.

Rubenfire, A. 2016. CMS and FDA advocate for device identifiers on claims forms. Available at http://www.modern-healthcare.com/article/20160714/NEWS/160719938 (accessed August 25, 2016).

Silberzahn, R., and E. L. Uhlmann. 2015. Crowdsourced research: Many hands make tight work. *Nature* 526(7572):189-191.

Smith, P., R. Araya-Guerra, C. Bublitz, B. Parnes, L.M. Dickinson, R. Van Vorst, J.M. Westfall, and W.D. Pace. 2005. Missing clinical information during primary care visits. *Journal of the American Medical Association* 293(5):565-571.

Straumsheim, C. 2014. Gates goes open. Available at https://www.insidehighered.com/news/2014/11/24/gates-foundation-announces-open-access-policy-all-grant-recipients (accessed August 25, 2016).

Taichman, D. B., J. Backus, C. Baethge, H. Bauchner, P. W. de Leeuw, J. M. Drazen, J. Fletcher, F. A. Frizelle, T. Groves, A. Haileamlak, A. James, C. Laine, L. Peiperl, A. Pinborg, P. Sahni, and S. Wu. 2016. Sharing clinical trial data: A proposal from the International Committee of Medical Journal Editors. *Annals of Internal Medicine* 164(7):505-506.

Tulsa Beacon Community. 2012. Available at https://www.healthit.gov/sites/default/files/beacon-fact-sheet-tulsa.pdf (accessed August 25, 2016).

Wellcome Trust. No date a. Data sharing webpage. Available at http://www.wellcome.ac.uk/About-us/Policy/Spotlight-is-sues/Data-sharing/ (accessed August 25, 2016).

Wellcome Trust. No date b. Why we're launching a new publishing platform. Available at https://wellcome.ac.uk/news/why-were-launching-new-publishing-platform (accessed August 25, 2016).

Wilhelm, E., E. Oster, and I. Shoulson. 2014. Approaches and costs for sharing clinical research data. *Journal of the American Medical Association* 311(12):1201-1202.

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, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. t Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S. A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons. 2016. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data* 3:160018.

Suggested Citation

Krumholz, H.M., P.E. Bourne, R.E. Kuntz, M. McClellan, H.L. Paz, S.F. Terry, and J. Waldstreicher,. 2016. *Data Acquisition, Curation, and Use for a Continuously Learning Health System.* Vital Directions for Health and Health Care Series. Discussion Paper, National Academy of Medicine, Washington, DC. http://nam.edu/wp-content/uploads/2016/09/data-acquisition-curation-and-use-for-a-continuously-learning-health-system.pdf

Author Information

Harlan M. Krumholz, MD, SM, is Harold H. Hines, Jr. Professor of Medicine and Epidemiology and Public Health, Yale University School of Medicine. **Philip E. Bourne, PhD,** is Associate Director for Data Science, National Institutes of Health. **Richard E. Kuntz, MD, MSc,** is Senior Vice President, Chief Scientific, Clinical and Regulatory Officer, Medtronic, Inc. **Harold L. Paz, MD, MS,** is Executive Vice President, Chief Medical Officer, Aetna. **Sharon F. Terry, MA,** is President and CEO, Genetic Alliance. **Joanne Waldstreicher, MD,** is Chief Medical Officer, Johnson & Johnson.

Acknowledgments

Sponsors

Disclaimer