

Perspective

People-powered data collaboratives: fueling data science with the health-related experiences of individuals

Barbara J Evans¹ and Harlan M Krumholz^{2,3,4}

¹Law Center and Department of Electrical and Computer Engineering, University of Houston, Houston, Texas, USA, ²Section of Cardiovascular Medicine, Department of Internal Medicine, Yale School of Medicine, New Haven, Connecticut, USA, ³Department of Health Policy and Management, Yale School of Public Health, New Haven, Connecticut, USA, and ⁴Center for Outcomes Research and Evaluation, Yale-New Haven Hospital, New Haven, Connecticut, USA

Corresponding Author: Harlan M Krumholz, MD, SM, Yale Center for Outcomes Research and Evaluation, 1 Church Street, Suite 200, New Haven, CT 06510, USA (harlan.krumholz@yale.edu)

Received 26 June 2018; Revised 25 October 2018; Editorial Decision 1 November 2018; Accepted 13 November 2018

ABSTRACT

The creation of people-driven data collaboratives, with governance structures that enable participants to have a meaningful voice in issues surrounding the use of their own data, is a novel strategy to harness our growing capacity to develop and maintain immense data assets from the real health experiences of individuals.

Key words: data collaboratives, information dissemination, data access, data citizenship

INTRODUCTION

A remarkable age of discovery could evolve from the growing capacity to acquire immense amounts of data from the real health experiences of individuals and from new opportunities to characterize their biology, behaviors, exposures, and outcomes. Transforming this prospect into reality depends on access to data, which are the vital reagents for work that can generate insights possible only through the study of massive amounts of health-related data from large numbers of people.

The challenge is to create the means for massive numbers to contribute their data to research. Two key questions are: Are traditional, 20th-century informed consent models scalable to the diverse, inclusive, multimillion-person data sets that 21st-century science requires? If not, are there ethically acceptable alternatives, and what might these look like?

A “traditional” consent model is an approach that empowers individuals to control access to data about themselves by consenting to proposed data uses. What distinguishes this model is its reliance on individual rather than collective decision making. Each person, acting alone, decides the fate of his or her own data. This article proposes a consent model that engages people as data citizens who decide the fate of their data—and the terms on which

they can be used—through a process of collaborative self-governance.

THE DATA ACCESS CHALLENGE

Data resources that capture people’s real health experiences must meet several basic expectations. Ideally, the data are aggregated with the knowledge and permission of those the data describe, respecting people’s strong desire to control uses of their data. Data remain secure and protected through all phases of transmission and use. All data sharing and uses comply with legal requirements, reflect state-of-the-art privacy and security policies, and enjoy the public’s trust. Data are readily transformed into common definitions to support analysis and can be combined seamlessly across vendors and data types. Data are frequently updated and appropriately linked to represent a complete longitudinal record of each individual. System governance is transparent and trustworthy so as to attract broad public engagement in the shared mission of creating and maintaining large, inclusive data resources.

These ideals, unfortunately, outpace present realities. Regulations such as the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule¹ and the Common Rule² offer an opt-in

consent scheme. People grant or deny consent to data requests initiated by others but have little proactive power to influence data uses or negotiate terms, such as the privacy protections they desire. Surveys show that up to 80% of Americans would like to contribute their data for socially beneficial research³ yet do not do so. There are likely many reasons that people choose not to contribute their data to research, but one possibility, among many, is that people may not find the standard privacy protections and regulatory terms to be suitable.

Absent individual consent, the regulations supply an array of consent exceptions that allow data holders to drive data access. With approval from an Institutional Review Board (IRB), data holders can use opt-out consent, often presented to patients when they are ill and dependent on the institution that is asking them to acquiesce. In addition, the HIPAA Privacy Rule has exceptions that let data holders share people's data without individual authorization if data are de-identified, disclosed to a public health authority, or disclosed for research pursuant to an IRB-approved waiver.⁴ The Common Rule has analogous provisions. These exceptions empower willing institutional data holders, acting on their own initiative, to drive data access.

These regulatory exceptions unleashed data-holder-driven strategies that became major workhorses of data access in recent decades. Each data holder is a single point of access to data for the many individuals from which it has generated data. Data crossing a data holder's privacy firewall in fully or partially de-identified form, however, cannot easily be linked with data from other sources to assemble longitudinal records for individuals. One work-around is to leave data distributed behind the data holders' privacy firewalls and create the means to perform analyses locally and combine them centrally. This approach circumvented the need to manage data use agreements among the data holders and avoided the need for them to share data. Unfortunately, distributed approaches constrain analyses in ways that limit their full potential to produce new knowledge. As a last resort, IRB-approved consent waivers can move data across firewalls in identifiable form, enabling creation of longitudinal records.

The looming failure of data-holder-driven access

These traditional access strategies face new challenges. Twenty-first-century science often requires multidimensional data assembly—not only from many individuals but also across many different data holders that store portions of each person's data—to capture people's complete experiences over time. De-identifying data can thwart necessary linkages, and there is growing awareness that large, linked datasets are inherently re-identifiable, which makes many IRBs reluctant to approve waivers. Data holders are efficient at aggregating data across the individuals included in their data sets, but data holders have no special advantage with regard to sharing data with each other or linking across diverse data sources. They face legal barriers, such as HIPAA's minimum necessary constraints on inter-institutional data sharing.⁵ There are many reasons that data holders may not share data, but data holders generally prize their data assets and may have commercial incentives not to share, in addition to bureaucratic burden, lack of understanding of the regulations, and lack of expertise in mechanisms of data sharing.

Institutional data holders regulated by HIPAA and the Common Rule have significant power to drive data access but lack a financially sustainable revenue model to cover the costs of converting data into interoperable formats and maintaining data resources for long-term use. HIPAA's cost-based fee for data preparation and transmittal lets them recover only part of these costs.⁶ Governmental incentives—such as conditioning grants on data holders' willingness

to share data—can advance discrete projects but are neither sustainable nor scalable without substantial continued funding.

Data-holder-driven strategies skirt the need for patient engagement, which may seem expedient but can ignore individuals' preferences about the use of their data. They also limit opportunities to integrate patient-reported data into a shared data resource. Nontraditional data holders, such as mobile device manufacturers, are subject to different regulations and hold valuable data that traditional regulatory access strategies may miss.⁷ Yet, overreliance on traditional clinical data may miss opportunities to compose data resources that give a unified, longitudinal representation of individuals' health-related experiences in and out of healthcare institutions.

Neither the individual nor the data holder has exclusive legal ownership of stored health information.^{7,8} Both parties have legitimate interests in the information and share control. Ideally, sharing would involve consent alignment, whereby individuals and the institutions that hold their data both consent to share data for research, but this is unlikely to emerge on the scale needed to create the massive data resources that science now requires. The government could compel access through laws that mandate data sharing, but legislators historically have invoked these powers sparingly.⁷

THE PROMISE OF PEOPLE-DRIVEN DATA COLLABORATIVES

A truly people-centered system would acknowledge the agency of individuals over their own data and harness the potential of partnering with people to assemble high-quality longitudinal data resources. This approach demands 2 elements. The first is a legally enforceable individual access right that empowers people to free their data from balky data holders. The second is people-centered consensual arrangements that foster their collaborative efforts to combine their data into high-valued collective data resources.

The Privacy Rule already grants people a right of access to their designated record set, which includes medical, billing, and other information held by HIPAA-covered entities in digital and physical files. This access expanded in 2014 to include data held by HIPAA-covered laboratories.⁹ Next steps are for the Department of Health and Human Services' Office for Civil Rights to aggressively enforce this right, and for bodies such as the Federal Trade Commission, Food and Drug Administration, and state legislatures to collaborate to ensure a similar right for consumers to access data held by non-HIPAA entities such as device manufacturers.

The larger challenge is to make it simple, attractive, and perhaps even enticing for people to work together to build large, collective data resources and to give people a meaningful voice in deciding how, and on what terms, their data can be used. We propose people-powered data collaboratives: self-governing communities of individuals empowered by access to their own data. These could be organized by the members, by advocacy groups, or by commercial data management companies that commit to administer each collaborative according to rules the members set. Such an arrangement allows for the power of collective data governance and, ultimately, collective wisdom. People join together to help those who will follow them—and, once in place, will know that they were helped by those before them.

The specific organizational structures of these collectives and the funding and governance arrangements could be established in a variety of ways, as has been described in some detail in the legal literature.^{7,10} New legislation would not be required because collective

organizations can be set up contractually. However, legislation might be beneficial in order to set up basic ground rules—such as a list of basic rights and protections people should have when contributing their data to a consumer-driven data commons—and to help create incentives for the creation of such commons.

Individuals could voluntarily join and contribute their data to 1 or more collaboratives (or none). The data could include clinical information, patient-generated data, and patient-reported data. People bringing together their own data could create a comprehensive, longitudinal, continuous data stream that describes their health journey and health experiences. In a digital age, such data transmission is possible through application program interfaces that can move data, with permission, from many sources and do so automatically.

Each collaborative would publish transparent rules describing how to join and exit, and explaining the members' rights and duties. Members would have a right to participate in the collaborative's self-governance process and vote on the types of research for which their collective data resources could be used and the consent, privacy, and data security standards they would require. For example, some collaboratives might vote to implement granular, opt-in, individual consent, while others vote to make their data assets more valuable by offering their entire, collective data resources for uses approved by a majority of the members. The individual's right of consent takes the form of consenting to participate in 1 or more collaborative groups, in which data uses would be decided by transparently disclosed rules set by the members. The idea is to ensure that people have agency over their data and to provide the means for them to come together as a collective—not controlled by other organizations—but with autonomy to make their own decisions. These organizations can then hire the expertise they need to address issues such as data security, privacy, and prioritizing various uses of their data.

The collaboratives could be constituted of people who suffer from a chronic condition or have undergone a particular procedure or received a certain drug or had a specific device implanted or have an interest in health. Like all studies that require consent (in this case, consent to membership in the group), the collaborative groups may not be fully representative of all people with these conditions, but over time it may become normative for people with these conditions to join others to enable medicine to progress by learning from their experiences.

The advantage of this approach is that it is entirely consent based. The relevant act of consent lies in joining, remaining in, or withdrawing from a collaborative—or in participating in certain activities of the collective. It also harnesses collective bargaining as a tool of human-subject protection. A vibrant marketplace of people-driven ethical and privacy standards could emerge as different collaboratives enunciate their terms of access and compete for members. People-driven collaboratives can propel and accelerate knowledge generation.

Importantly, people-driven collaboratives offer the prospect of sustainable revenue models to cover the costs of data sharing and ongoing operation. There is no legal restriction on people's ability to assess fees for the use of their own data. Collaboratives whose members vote to do so could levy user fees to cover the costs of hiring skilled technical support and making investments to enhance the value of their collective data assets. They also could devise pecuniary and non-financial incentives to attract members and reward data sharing.

CONCLUSION

Data resources that capture the real health experiences of large numbers of people are the fuel for 21st-century discovery but will

require bold steps to overcome cultural and regulatory barriers to data sharing. It is time to position people, rather than institutional data holders, as central drivers of data access. They could be the means by which diverse data sources are aggregated and organized, in and out of clinical encounters. The approaches outlined here are envisioned as complementary to other research approaches, transforming what is possible in a digital age. People-driven collaboratives are a bold approach to engage people, overcome regulatory barriers, and create data resources that reflect the public's real health experiences.

DISCLOSURES

Harlan Krumholz is a recipient of research agreements from Medtronic and from Johnson & Johnson (Janssen), through Yale, to develop methods of clinical trial data sharing; was the recipient of a research grant from Medtronic and the Food and Drug Administration, through Yale, to develop methods for post-market surveillance of medical devices; works under contract with the Centers for Medicare & Medicaid Services to develop and maintain performance measures that are publicly reported; chairs a cardiac scientific advisory board for UnitedHealth; is a participant/participant representative of the IBM Watson Health Life Sciences Board; is a member of the Advisory Board for Element Science and the Physician Advisory Board for Aetna; and is the founder of Hugo, a personal health information platform. Barbara Evans has no conflicts to disclose.

FUNDING

Barbara Evans received funds from NIH/NHGRI/NIC R01 HG008605, U01 HG006507, U01 HG007307 and the Robert Wood Johnson Foundation Health Data Exploration Project. Views expressed are those of the author rather than the views of the funders.

CONTRIBUTORS

Drs Evans and Krumholz wrote the manuscript, revised it for important intellectual content, and approved the final for submission.

Conflict of interest statement. None declared.

REFERENCES

1. Health Insurance Portability and Accountability Act of 1996 ("HIPAA," Pub. L. No. 104-191, 1996) Privacy Rule, 45 C.F.R. pts. 160, 164.
2. 45 C.F.R. pt. 46, subpt. A.
3. Kish LJ, Topol EJ. Unpatients—why patients should own their medical data. *Nat Biotechnol* 2015; 11: 921–94.
4. 45 C.F.R. §§ 164.502(d)(2), 512(b)(1)(i), 512(i).
5. 164.502(b), 164.514(d).
6. Evans BJ. Mining the human genome after Association for Molecular Pathology v. Myriad Genetics. *Genet Med* 2014; 16 (7): 504–9.
7. Evans BJ. Barbarians at the gate: consumer-driven data commons and the transformation of citizen science. *Am J Law Med* 2016; 42 (4): 651–85.
8. Organization for Economic Cooperation and Development. Data-driven innovation: big data for growth and wellbeing. *OECD* 2015; doi: 10.1787/9789264229358-en.
9. 45 C.F.R. § 164.524.
10. Evans BJ. Power to the people: data citizens in the age of precision medicine. *Vanderbilt J Entertain Technol Law* 2017; 19 (2): 243–65.