A distributed health data network is a system that allows secure remote analysis of separate data sets, each derived from a different medical organization’s or health plan’s records. Such networks allow data holders to retain physical control over use of their data, thereby avoiding many obstacles related to confidentiality, regulatory, and proprietary interests. They can be used for observational studies, particularly public health surveillance, and can also provide baseline and follow-up data to support clinical trials, including those that use cluster randomization. In addition, a network can monitor use, adoption, and diffusion of new technologies and clinical evidence. Such networks are critical elements of the “learning health care system” recommended by the Institute of Medicine (1), which supports the use of routinely collected health care data to improve our understanding of the comparative benefits and harms of medical technologies.

The United States will soon be able to analyze data from millions of individuals. Congress has mandated that the U.S. Food and Drug Administration develop a post-market risk identification and analysis system that covers 100 million persons (2). In addition, the expansion of comparative effectiveness research envisioned by Congress requires access to health care information for large, diverse populations in real-world settings (3). Large, centralized data repositories could support these functions, but we and others (4, 5) believe that a distributed health data network has many practical advantages. First, a distributed network allows data holders to retain physical and logical control of their data. Second, it mitigates many security, proprietary, legal, and privacy concerns, including those regulated by the Privacy and Security Rules of the Health Insurance Portability and Accountability Act (6). Third, it eliminates the need to create, maintain, and secure access to central data repositories. Fourth, it minimizes the need to disclose protected health information outside the data-owning entity. Finally, a distributed network allows data holders to assess, track, and authorize requests for all data uses.

Several public agencies have supported the development of single-purpose distributed data networks, either directly or in principle (7–11). These networks are limited in scope and do not support the broad range of public and private needs filled by the network we describe. We favor a single distributed network with multiple uses—for example, one that could be used to study comparative clinical effectiveness and the diffusion of medical technologies—over multiple independent and single-purpose networks. A multipurpose network would reduce the burden on data holders of participating in multiple networks, as well as that on network developers of creating and maintaining redundant infrastructure. The framework that we describe suggests how we could develop a national network with broad capabilities.

HOW WOULD A NATIONAL DISTRIBUTED HEALTH DATA NETWORK WORK?

In the simplest national distributed health data network, each data holder creates a copy of their data (a “network datamart”) that adheres to a common data model, thus ensuring identical file structures, data fields, and coding systems. Several common data models already exist (10, 12–17). The Figure illustrates the basic flow of network operations. Authorized users submit queries by means of a secure Web site. Data holders set authorization policies for each user and query type and can require approvals from privacy boards and institutional review boards. The network interface allows nontechnical users to ask simple questions without assistance (for example, a report on the uptake of a given treatment by age, sex, and geographic region). It also allows sophisticated users to perform com-
plex analyses (for example, comparing the rates of serious cardiovascular outcomes among patients who receive different second-line antihypertensive treatments). For many questions, transferring protected health information will not be necessary. However, it may be necessary to aggregate relatively small amounts of data for analysis. Using the network, data holders may provide limited access to full-text medical records for validation and additional details. It is usually necessary to review only a small proportion of records to confirm diagnoses or to obtain risk factor data that are not coded (such as smoking status).

**Example of the Use of a Distributed Network**

Some research programs already use a distributed network model (10, 14, 18), which provides a relevant starting point to implement a national network. The HMO Research Network Center for Education and Research on Therapeutics has conducted many multisite studies by distributing computer programs that each site applied to a local copy of their data. The outputs are then combined to provide aggregate results. Examples of studies performed in this way include the evaluation of laboratory monitoring practices for medications (18–25), the use of medications during pregnancy (26–28), and the use of medications that carry a black box warning (29). Such studies provide an important evidence development function that feeds back to providers, payers, and patients.

**Policy Issues**

Development and implementation of a multipurpose, multi-institutional distributed health data network requires substantial stakeholder engagement and dedicated software development. On the basis of the previously described research studies, we recommend incremental implementation with a limited set of data holders and data types. Begin with information about eligibility for health care (such as health plan enrollment data); this would allow identification of defined populations, which are important for many uses. Initial data should also include demographic characteristics; diagnosis, procedure, and pharmacy dispensing data (30); and, potentially, electronic health record data, such as vital signs. During initial implementation, pilot testing is needed to assess network design, software development, and development and implementation of the common data model.

A distributed network’s viability depends on both its governance mechanisms and sustained funding. A governance institution is needed to develop and oversee processes.
Design of a National Distributed Health Data Network

Academia and Clinic

Condures for requesting use of the network; to set priorities; and to audit use for compliance with various security, privacy, human subject research, and proprietary concerns. Such an institution should also monitor research integrity, data integrity, conflict of interest policies, transparency of activity and results, policies related to access and use, reproducibility, publishing rights, and dispute resolution.

Annual development and maintenance costs would probably be several tens of millions of dollars for an initial system that covers up to 100 million persons. This would be similar to the 3-year startup cost for the National Cancer Institute’s Cancer Biomedical Informatics Grid, which totaled $60 million for fiscal years 2004 to 2006 (31). The National Cancer Institute fiscal year 2010 budget requests $100 million for these efforts in addition to the current funding level (32). The total annual cost of developing and maintaining a network is in line with that of individual clinical trials routinely performed to evaluate new pharmaceuticals. Although initial implementation costs are sizeable, the expected marginal costs to use the system would be small for any particular study. Various funding mechanisms are possible. Initially, we expect costs to be borne by the federal entities, whose current needs would drive network implementation. Ultimately, we believe the costs should be amortized over the system’s multiple users and should support the network’s expansion, functionality, and use. For example, methods could be developed for linking to the National Death Index or identifying individuals for whom multiple data holders possess different kinds of information (such as pharmacy data held by one source and clinical encounter data held by another). Advances in technologies designed to link individual records over time (such as anonymous identity resolution) without exposing protected health information are especially desirable (33).

Conclusion

A national distributed health data network can become an important asset to improving health and health care. A common core network would offer considerable advantages that would better support the needs of multiple users, such as the U.S. Food and Drug Administration (for their Sentinel System) and the Agency for Healthcare Research and Quality (for their comparative effectiveness network), than would building individual networks for each of these uses. The similarities in data needs and uses, coupled with potential savings of time and effort, favor a single, multipurpose network. In addition, local data holder control over use and access would encourage participation. Finally, credible governance and funding mechanisms are critical to ensure the long-term sustainability of the network. Development of a multipurpose, multi-institutional distributed health data network would accelerate the development of a learning health care system.

From Massachusetts Institute of Technology, Cambridge, and Harvard Medical School and Harvard Pilgrim Health Care, Boston, Massachusetts; and University of Pennsylvania, Philadelphia, Pennsylvania.


Disclaimer: The authors of this report are responsible for its content. Statements in this report should not be construed as endorsement by the Agency for Healthcare Research and Quality or the U.S. Department of Health and Human Services.

Acknowledgment: The authors thank Kimberly Lane, MPH, and Beth Syat, MPH, for their project support and Scott Smith, PhD, for his helpful comments.

Grant Support: By the Agency for Healthcare Research and Quality, U.S. Department of Health and Human Services (contract no. 290-05-0033), as part of the Developing Evidence to Inform Decisions about Effectiveness (DEcIDE) program.


Requests for Single Reprints: Jeffrey S. Brown, PhD, Department of Ambulatory Care and Prevention, Harvard Medical School and Harvard Pilgrim Health Care, 133 Brookline Avenue, 6th Floor, Boston, MA 02215; e-mail, jeff_brown@harvardpilgrim.org.

Current author addresses are available at www.annals.org.

1 September 2009 | Annals of Internal Medicine | Volume 151 • Number 5 | 343
References

Current Author Addresses: Ms. Maro: Engineering Systems Division, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Building 41-205, Cambridge, MA 02139.
Drs. Platt and Brown: Department of Ambulatory Care and Prevention, Harvard Medical School and Harvard Pilgrim Health Care, 133 Brookline Avenue, 6th Floor, Boston, MA 02215.
Dr. Holmes: Center for Clinical Epidemiology and Biostatistics, University of Pennsylvania School of Medicine, 726 Blockley Hall, 423 Guardian Drive, Philadelphia, PA 19104-6021.
Dr. Strom: Center for Clinical Epidemiology and Biostatistics, University of Pennsylvania School of Medicine, 824 Blockley Hall, 423 Guardian Drive, Philadelphia, PA 19104-6021.
Dr. Hennessy: Center for Clinical Epidemiology and Biostatistics, University of Pennsylvania School of Medicine, 803 Blockley Hall, 423 Guardian Drive, Philadelphia, PA 19104-6021.
Mr. Lazarus: Channing Laboratory, 181 Longwood Avenue, Boston, MA 02115.