Design of a National Distributed Health Data Network

Judith C. Maro, MS; Richard Platt, MD, MSc; John H. Holmes, PhD; Brian L. Strom, MD, MPH; Sean Hennessy, PharmD, PhD; Ross Lazarus, MBBS, MPH; and Jeffrey S. Brown, PhD

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. Distributed health data networks are currently being planned that could cover millions of people, permitting studies of comparative clinical effectiveness, best practices, diffusion of medical technologies, and quality of care. These networks could also support assessment of medical product safety and other public health needs. Distributed network care. These networks could also support assessment of medical product safety and other public health needs. Distributed network technologies allow data holders to control all uses of their data, which overcomes many practical obstacles related to confidentiality, regulation, and proprietary interests. Some of the challenges and potential methods of operation of a multipurpose, multi-institutional distributed health data network are described.

For author affiliations, see end of text.
This article was published at www.annals.org on 28 July 2009.

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-

See also:
Print
Key Summary Points ............................................. 342
Editorial comment .................................................. 359
Related article ....................................................... 338
Web-Only
Conversion of graphics into slides

© 2009 American College of Physicians
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 process.
Design of a National Distributed Health Data Network

Academia and Clinic

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.
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.