

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 comprising 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 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.

Ann Intern Med. 2009;151.
For author affiliations, see end of text.
This article was published at www.annals.org on 28 July 2009.

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, regulation, 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 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 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-

See also:

Print

Key Summary Points.	1
Editorial comment.	2
Related article.	3

Web-Only

Conversion of graphics into slides

Key Summary Points: Attributes of a National Distributed Health Data Network

Supports both observational and intervention studies.

Local data holder control over access and uses of data.

Mitigates need to share or exchange protected health information.

Singular, multipurpose, multi-institutional infrastructure.

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 fraction 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

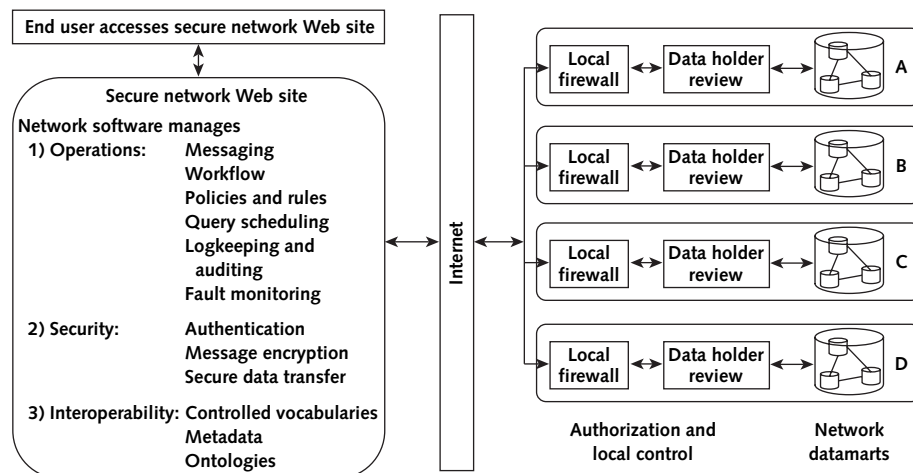
Therapeutics has conducted many multisite studies by distributing computer programs that each site applied to a local copy of their data. The center then combined the outputs to provide aggregate results. Examples of studies performed in this way include the evaluation of laboratory monitoring practices 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 proce-

Figure. System operations.



An authorized user accesses the secure network Web site to submit queries (computer programs) to run against data in the network datamarts. The solid boxes depict areas under control of the data holder (data holders A–D are shown). Authorization to execute a query is under control of the data holder and can be limited to specific users and uses. Data holders retrieve queries for execution, which eliminates the need for data holders to monitor incoming requests. Query results are encrypted and returned to the central Web site where they are processed and presented to the requestor. Details of each step are recorded for auditing.

dures 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 covered 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 for 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, Massachusetts; Harvard Medical School and Harvard Pilgrim Health Care, Boston, Massachusetts; and University of Pennsylvania, Philadelphia, Pennsylvania.

Note: Several ideas expressed here were presented at an Institute of Medicine workshop in December 2007 and discussed in a follow-on publication: **Platt R.** Distributed data networks. In: Institute of Medicine. Redesigning the Clinical Effectiveness Paradigm—Innovation and Practice based Approaches. Washington, DC: Nat Acad Pr; 2009.

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.

Potential Financial Conflicts of Interest: *Consultancies:* B.L. Strom (Abbott Laboratories, Bayer Corporation, Bristol-Myers-Squibb, Daichii Pharmaceuticals UK, GlaxoSmithKline, Johnson & Johnson, Mediwound, Novartis Farmaceutica S.A., NPS Pharmaceuticals, Oscient, Pfizer, Sanofi Aventis, Teva Neuroscience, Wyeth), S. Hennessy (Wyeth; Teva; and law firms representing Bayer, Eli Lilly, and Pfizer), J.S. Brown (Phase Forward). *Honoraria:* J.H. Holmes (American Medical Informatics Association), B.L. Strom (Johnson & Johnson, Washington University in St. Louis), J.S. Brown (HealthCore). *Expert testimony:* R. Platt (Agency for Healthcare Research and Quality, U.S. Food and Drug Administration). *Grants received:* R. Platt (Sanofi-Aventis, GlaxoSmithKline, Pfizer, TAP Pharmaceuticals, Agency for Healthcare Research and Quality, Centers for Disease Control and Prevention, U.S. Food and Drug Administration, National Institutes of Health, America's Health Insurance Plans, Massachusetts Department of Public Health), J.H. Holmes (Agency for Healthcare Research and Quality, National Library of Medicine, National Cancer Institute, National Institute for Allergy and Infectious Disease, National Institute for Diabetes and Digestive and Kidney Diseases), B.L. Strom (National Institutes of Health, Agency for Healthcare Research and Quality, Takeda Pharmaceuticals, Shire Development), S. Hennessy (Agency for Healthcare Research and Quality, National Institutes of Health, Shire Development), J.S. Brown (Sanofi-Aventis, GlaxoSmithKline, and Pfizer, Agency for Healthcare Research and Quality, Centers for Disease Control and Prevention, U.S. Food and Drug Administration, National Institutes of Health). *Grants pending:* J.H. Holmes (National Library of Medicine, National Cancer Institute, National Institute for Neurological Disorders and Stroke), B.L. Strom (National Institutes of Health, AstraZeneca, Bristol-Meyers Squibb), S. Hennessy (U.S. Food and Drug Administration).

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

1. Institute of Medicine. Learning Healthcare System Concepts v2008: Annual Report. Washington, DC: Nat Acad Pr; 2008.
2. Public Law 110-85. Food and Drug Administration Amendments Act of 2007. 121 STAT. 823.
3. Public Law 111-5. Title VIII of American Recovery and Reinvestment Act of 2009. 123 STAT. 115.
4. U.S. Food and Drug Administration, eHealth Initiative Foundation, Brookings Institute. Sentinel Initiative: Structure, Function, and Scope [Transcript]. Public Meeting. 16 December 2008; Washington, DC. Accessed at www.fda.gov/downloads/ScienceResearch/SpecialTopics/CriticalPathInitiative/PastEventsonCPI/UCM113469.pdf on 17 July 2009.
5. Diamond CC, Mostashari F, Shirkly C. Collecting and sharing data for population health: a new paradigm. *Health Aff (Millwood)*. 2009;28:454-66. [PMID: 19276005]
6. Moore KM, Duddy A, Braun MM, Platt R, Brown JS. Potential population-based electronic data sources for rapid pandemic influenza vaccine adverse event detection: a survey of health plans. *Pharmacoepidemiol Drug Saf*. 2008;17:1137-41. [PMID: 18763248]
7. U.S. Department of Health and Human Services, U.S. Food and Drug Administration. The Sentinel Initiative: A National Strategy for Monitoring Medical Product Safety. Rockville, MD: U.S. Food and Drug Administration; 2008. Accessed at www.fda.gov/Safety/FDAsSentinelInitiative/ucm089474.htm on 17 July 2009.
8. Agency for Healthcare Research and Quality. Developing a Distributed Research Network to Conduct Population-based Studies and Safety. Rockville, MD: Agency for Healthcare Research and Quality; 2008. Accessed at www.effectivehealthcare.ahrq.gov/healthInfo.cfm?infotype=nr&ProcessID=54 on 17 July 2009.
9. Agency for Healthcare Research and Quality. Distributed Network for Ambulatory Research in Therapeutics. Rockville, MD: Agency for Healthcare Research and Quality; January 2008. Accessed at www.effectivehealthcare.ahrq.gov/healthInfo.cfm?infotype=nr&ProcessID=53 on 15 August 2008.
10. Platt R, Davis R, Finkelstein J, Go AS, Gurwitz JH, Roblin D, et al. Multicenter epidemiologic and health services research on therapeutics in the HMO Research Network Center for Education and Research on Therapeutics. *Pharmacoepidemiol Drug Saf*. 2001;10:373-7. [PMID: 11802579]
11. Halamka J, Aranow M, Ascenzo C, Bates D, Debor G, Glaser J, et al. Health care IT collaboration in Massachusetts: the experience of creating regional connectivity. *J Am Med Inform Assoc*. 2005;12:596-601. [PMID: 16049225]
12. McMurry AJ, Gilbert CA, Reis BY, Chueh HC, Kohane IS, Mandl KD. A self-scaling, distributed information architecture for public health, research, and clinical care. *J Am Med Inform Assoc*. 2007;14:527-33. [PMID: 17460129]
13. Lazarus R, Yih K, Platt R. Distributed data processing for public health surveillance. *BMC Public Health*. 2006;6:235. [PMID: 16984658]
14. Chen RT, Glasser JW, Rhodes PH, Davis RL, Barlow WE, Thompson RS, et al. Vaccine Safety Datalink project: a new tool for improving vaccine safety monitoring in the United States. The Vaccine Safety Datalink Team. *Pediatrics*. 1997;99:765-73. [PMID: 9164767]
15. Yih WK, Caldwell B, Harmon R, Kleinman K, Lazarus R, Nelson A, et al. National Bioterrorism Syndromic Surveillance Demonstration Program. *MMWR Morb Mortal Wkly Rep*. 2004;53 Suppl:43-9. [PMID: 15714626]
16. Hornbrook MC, Hart G, Ellis JL, Bachman DJ, Ansell G, Greene SM, et al. Building a virtual cancer research organization. *J Natl Cancer Inst Monogr*. 2005;12-25. [PMID: 16287881]
17. Observational Medical Outcomes Partnership. Observational Medical Outcomes Partnership: Points to Consider in Developing a Common Semantic Data Model and Terminology Dictionary for Observational Analyses. Rockville, MD: Foundation for the National Institutes of Health; 2009. Accessed at http://omop.fnih.org/sites/default/files/OMOP%20Points%20to%20Consider%20on%20Common%20Data%20Model%203mar2009_Post.pdf on 17 July 2009.
18. Chan KA; HMO Research Network CERT Patient Safety Study Investigators. Development of a multipurpose dataset to evaluate potential medication errors in ambulatory settings. In: Henriksen K, Battles JB, Marks ES, Lewin DI, eds. *Advances in Patient Safety: from Research to Implementation*. Vol. 2: Concepts and Methodology. AHRQ Publication No. 05-0021-2. Rockville, MD: Agency for Healthcare Research and Quality; 2005:225-38.
19. Raebel MA, McClure DL, Chan KA, Simon SR, Feldstein AC, Lafata JE, et al. Laboratory evaluation of potassium and creatinine among ambulatory patients prescribed spironolactone: are we monitoring for hyperkalemia? *Ann Pharmacother*. 2007;41:193-200. [PMID: 17284506]
20. Raebel MA, Carroll NM, Simon SR, Andrade SE, Feldstein AC, Lafata JE, et al. Liver and thyroid monitoring in ambulatory patients prescribed amiodarone in 10 HMOs. *J Manag Care Pharm*. 2006;12:656-64. [PMID: 17269844]
21. Raebel MA, Carroll NM, Andrade SE, Chester EA, Lafata JE, Feldstein A, et al. Monitoring of drugs with a narrow therapeutic range in ambulatory care. *Am J Manag Care*. 2006;12:268-74. [PMID: 16686584]
22. Raebel MA, McClure DL, Simon SR, Chan KA, Feldstein AC, Gunter MJ, et al. Frequency of serum creatinine monitoring during allopurinol therapy in ambulatory patients. *Ann Pharmacother*. 2006;40:386-91. [PMID: 16478808]
23. Raebel MA, McClure DL, Simon SR, Chan KA, Feldstein A, Andrade SE, et al. Laboratory monitoring of potassium and creatinine in ambulatory patients receiving angiotensin converting enzyme inhibitors and angiotensin receptor blockers. *Pharmacoepidemiol Drug Saf*. 2007;16:55-64. [PMID: 16470693]
24. Raebel MA, Lyons EE, Andrade SE, Chan KA, Chester EA, Davis RL, et al. Laboratory monitoring of drugs at initiation of therapy in ambulatory care. *J Gen Intern Med*. 2005;20:1120-6. [PMID: 16423101]
25. Simon SR, Andrade SE, Ellis JL, Nelson WW, Gurwitz JH, Lafata JE, et al. Baseline laboratory monitoring of cardiovascular medications in elderly health maintenance organization enrollees. *J Am Geriatr Soc*. 2005;53:2165-9. [PMID: 16398903]
26. Andrade SE, Raebel MA, Brown J, Lane K, Livingston J, Boudreau D, et al. Outpatient use of cardiovascular drugs during pregnancy. *Pharmacoepidemiol Drug Saf*. 2008;17:240-7. [PMID: 18200619]
27. Andrade SE, Raebel MA, Brown J, Lane K, Livingston J, Boudreau D, et al. Use of antidepressant medications during pregnancy: a multisite study. *Am J Obstet Gynecol*. 2008;198:194.e1-5. [PMID: 17905176]
28. Andrade SE, Raebel MA, Morse AN, Davis RL, Chan KA, Finkelstein JA, et al. Use of prescription medications with a potential for fetal harm among pregnant women. *Pharmacoepidemiol Drug Saf*. 2006;15:546-54. [PMID: 16586470]
29. Wagner AK, Chan KA, Dashevsky I, Raebel MA, Andrade SE, Lafata JE, et al. FDA drug prescribing warnings: is the black box half empty or half full? *Pharmacoepidemiol Drug Saf*. 2006;15:369-86. [PMID: 16294363]
30. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. *J Clin Epidemiol*. 2005;58:323-37. [PMID: 15862718]
31. National Cancer Institute. Cancer Biomedical Informatics Grid Fact Sheet. Bethesda, MD: National Institutes of Health and U.S. Department of Health and Human Services; June 2006. Accessed at http://gforge.nci.nih.gov/docman/view.php/70/2709/caBIG_Fact_Sheet.pdf on 23 April 2009.
32. National Cancer Institute. The Nation's Investment in Cancer Research: Connecting the Nation's Cancer Community: An Annual Plan and Budget Proposal for Fiscal Year 2010. Bethesda, MD: National Institutes of Health and U.S. Department of Health and Human Services; 2009. Accessed at http://plan.cancer.gov/pdf/nci_2010_plan.pdf on 23 April 2009.
33. Jonas J. Identity resolution: 23 years of practical experience and observations at scale. In: Anonymous Proceedings of the 2006 ACM SIGMOD International Conference on Management of data. 27-29 June 2006; Chicago. New York: ACM; 2006:718.

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.

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.