Data Sharing and Embedded Research

Gregory E. Simon, MD, MPH; Gloria Coronado, PhD; Lynn L. DeBar, PhD, MPH; Laura M. Dember, MD; Beverly B. Green, MD, MPH; Susan S. Huang, MD, MPH; Jeffrey G. Jarvik, MD, MPH; Vincent Mor, PhD; Joakim Ramsberg, PhD; Edward J. Septimus, MD; Karen L. Staman, MS; Miguel A. Vazquez, MD; William M. Vollmer, PhD; Douglas Zatzick, MD; Adrian F. Hernandez, MD, MHS; and Richard Platt MD, MS

Sharing of data from clinical trials has the potential to increase transparency and reproducibility in medical research, enable secondary analyses, decrease selective reporting, and accelerate translation of high quality evidence into clinical care (1–3). Several solutions have been proposed to encourage the sharing of analyzable research data sets (4 – 8); however, the conceptual framework is rooted in explanatory clinical trials, which typically obtain explicit informed consent from participants and collect research-specific data focused on a narrow range of outcomes. Pragmatic research embedded in health systems often involves different data sources and data collection methods: It often involves a waiver of patient consent; uses data from the electronic health record; and may include information that could identify patients, health care providers, and health care facilities or organizations. Even if study data would not allow identification of individual participants, the potential for disclosure of sensitive information regarding providers or health systems may be substantial.

Although we enthusiastically support data sharing, potentially identifiable data regarding health systems or providers have the capacity to do harm if taken out of context; used for inappropriate comparisons; or used to single out individuals, providers, or institutions. Health care systems voluntarily participate in embedded research and have raised concerns about releasing unrestricted information from electronic health records. Specifically, health systems or facilities volunteering to participate in research might be penalized by release of detailed operational information that competitors are not required to make public. Measures developed for research may differ from publicly reported quality measures.

In an ideal world of transparency regarding health care processes and outcomes, health systems would have no expectation of or need for privacy regarding quality of health care delivery. However, the world is not perfect, and unintentional disclosure from participation in embedded research may be far greater than that required for public quality measures. Health systems volunteering to participate in research to improve public health may not be willing to bear the additional risk for misuse of sensitive information.

To encourage *individuals* to participate in clinical research, researchers offer explicit guarantees through the informed consent process that sensitive information will be protected and ensure that participants' protected health information is not exposed through trial activities or data sharing. Even when research studies are granted a waiver of consent to use patient information, researchers are bound to protect personal health information from disclosure. Although no such regulatory protection is in place for providers, practices, and health systems participating in research, a reasonable corollary exists. Such protections are especially important for providers included in cluster-randomized trials, in which explicit provider consent is uncommon. The notion that health systems, providers, or individual practitioners may be *participants* in embedded research—much like patients—has led some to argue for an ethical obligation to protect the privacy of health care providers and facilities. However, this ethical argument has proved contentious, especially given increasing expectations—or requirements—for transparency by hospitals, health systems, and the pharmaceutical and device industries.

Ultimately, the argument for protecting the privacy of health care systems and providers participating in research is a practical one. If those who volunteer to participate in research are required to bear significant additional risk, fewer will volunteer.

To motivate organizations to opt into embedded research for the greater good, we must recognize that sharing patient data might reveal sensitive information about providers or health systems. We recommend coupling that recognition with a framework for data sharing that champions making as much of the data available as possible for general use; allows additional analyses that refine or deepen the original research question, such as subsets or secondary outcomes; and encourages organizations to give serious consideration to other proposed uses while reserving the final authority regarding these decisions.

Researchers can assess risks by considering the sensitivity of each research data element and the risk that providers or facilities can be reidentified, and then reduce the risk by either modifying the data to be shared (such as redacting or masking sensitive data elements) or establishing governance structures appropriate to the level of risk. Potential structures for data sharing (ranging from least to most restrictive) include the following:

Public archive: Any interested users may download and analyze data without restriction. Private archive: Approved users may download and analyze data, sometimes subject to restrictions, often operationalized in a data use agreement.

Public enclave: Any interested users may submit queries and receive aggregate results. Private enclave: Approved users may submit queries and receive aggregate results (often subject to review and approval of individual queries).

Table. Data-Sharing Solutions Using Different Mechanisms*

Trial	ClinicalTrials.gov Identifier	Goal	Risks to Providers or Health Systems	What Data Can Be Shared Using a Public Archive Open to Any User?	Would a More Restrictive Structure Allow Sharing Additional Data of Substantial Public Health Value?	Data-Sharing Solution
PPACT	NCT02113592	Evaluate collaborative care program to improve self-management skills for chronic pain and limit use of opioid medications in primary care. Includes 830 patients in 3 health systems.	Naive comparisons of opioid prescribing rates across providers or facilities	Public data set will not include facility or health system identifiers or patient-level variables likely to allow reidentification of providers or facilities.	No. The primary analysis can still be closely replicated without the additional data.	Public archive of a modified data set
SPOT	NCT02326883	Evaluate population-based outreach programs to prevent suicide attempts in high-risk outpatients. Will include 16 000 patients in 4 health systems.	Naive comparisons of suicide attempts or suicide mortality across health systems	Public data set will not include an indicator for health system.	Possibly. Data sets, including health system identifiers, will be available on request through a supervised data archive, subject to formal agreements about use and redisclosure.	Public archive of a modified data set
STOP CRC	NCT01742065	Evaluate systematic EHR-embedded program for mailing of fecal immunochemical test kits to increase colorectal cancer screening rates in FQHCs. Will include 41 000 patients in 26 FQHC clinics.	Naive comparisons of cancer screening rates across providers or clinics. Perceived discrepancies between study outcome measures and publicly reported quality measures.	None	Yes. Deidentified patient-level data will be available on request, via a supervised data archive, subject to formal agreements about use, redisclosure, and data destruction.	Private archive managed by study team
ICD-Pieces	NCT02587936	Evaluate a novel technology platform (Pieces) supported by practice facilitation to improve care for patients with chronic kidney disease, diabetes, and hypertension within primary care practices or community medical homes. Will include 11 000 patients in 4 health systems.	Naive comparisons of care processes and chronic illness outcomes across providers or clinics. Perceived discrepancies between study outcome measures and publicly reported quality measures.	None	Yes. Patient-level data will be available via the NIDDK supervised data archive. Identifiers for health care system, primary practice, and patients will be removed.	Private archive managed by NIDDK
ABATE	NCT02063867	Evaluate the effect of daily antiseptic bathing, supplemented by nasal ointments for patients harboring MRSA, on clinical isolates of multidrug-resistant organisms and bloodstream infections attributable to non-critical care hospital units. Will include >600 000 patients in 53 hospitals.	Naive comparison of infection or complication rates across facilities. Perceived discrepancies between study outcome measures and publicly reported quality measures. Disclosure of proprietary business information about utilization patterns.	None	Yes. All individual-level data will remain behind the health system firewall. A private data enclave structure will allow potential users to propose specific queries, and only query results will be shared.	Private enclave managed by study team
LIRE	NCT02015455	Determine whether inserting epidemiologic benchmarks (essentially representing the normal range) into lumbar spine imaging reports reduces subsequent tests and treatments. Will include 4 health systems, 100 clinics, and 245 000 patients.	Naive comparison of opioid prescribing rates and other treatments across primary care providers and health systems. Disclosure of proprietary business information about utilization patterns.	None	Yes. Patient-level data sets will be deidentified by health systems, clinics, providers, and patients. Investigators will authorize release to specific users for specific purposes.	Private archive managed by study team

ABATE = Active Bathing to Eliminate Infection; EHR = electronic health record; FQHC = federally qualified health center; ICD-Pieces = Improving Chronic Disease Management with Pieces; LIRE = Lumbar Image Reporting with Epidemiology; MRSA = methicillin-resistant *Staphylococcus aureus*; NIDDK = National Institute of Diabetes and Digestive and Kidney Diseases; Pieces = Parkland Intelligent e-Coordination and Evaluation System; PPACT = Pain Program for Active Coping and Training; SPOT = Suicide Prevention Outreach Trial; STOP CRC = Strategies and Opportunities to Stop Colorectal Cancer in Priority Populations.
* Assumes Health Insurance Portability and Accountability Act of 1996–compliant patient deidentification for all patients and a data use agreement where appropriate.

A health care organization might allow partial data release by using less restrictive methods while requiring more restrictive methods for data it considers most sensitive.

More restrictive data-sharing structures necessarily require greater resources. Compared with a public archive, establishing a private archive requires personnel resources to review and approve users and specific uses. Compared with a data archive, establishing a data enclave to respond to users' queries requires substantially greater technical resources. When selecting an optimal technical and governance model for data sharing, investigators and participating health systems or practices should consider whether more restrictive (and expensive) approaches would allow sharing of additional data with significant added public health value. We recommend that the following questions be considered:

What data could be shared by the least restrictive mechanism, that is, a public archive open to any interested user?

What additional data could be shared by using a more restrictive mechanism (private archive, public or private data enclave)?

Would the scientific or public health benefit of sharing additional data justify the additional effort to establish a more restrictive data-sharing mechanism?

The research teams for selected demonstration projects of the Health Care Systems Research Collaboratory of the National Institutes of Health (NIH) were asked to consider these questions when creating a plan for sharing study data. The Table illustrates the solutions put forth by the teams.

We are confident that we can establish data sharing policies that will not dissuade health system participation. To balance potential for harm with the ethical imperative to share data, study teams can partner with health care systems to develop data-sharing plans that are the least restrictive and provide appropriate protection for participant privacy, health system privacy, and scientific integrity.

From Kaiser Permanente Washington Health Research Institute, University of Washington, and University of Washington School of Medicine, Seattle, Washington; Kaiser Permanente Center for Health Research, Portland, Oregon; University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania; University of California, Irvine School of Medicine, Orange, California; Brown University, Providence, Rhode Island; Harvard Pilgrim Health Care Institute and Harvard Medical School, Boston, Massachusetts; Hospital Corporation of America, Nashville, Tennessee; Texas A&M College of Medicine, Houston, Texas; CHB Wordsmith, Raleigh, North Carolina; University of Texas Southwestern Medical Center, Dallas, Texas; and Duke University School of Medicine, Durham, North Carolina.

Disclaimer: The views presented here are solely the responsibility of the authors and do not necessarily represent the official views of the NIH, the U.S. Department of Health and Human Services, or any of its agencies.

Grant Support: By a cooperative agreement (U54 AT007748) from the NIH Common Fund for the Coordinating Center of the NIH Health Care Systems Research Collaboratory and by the following NIH grants for the pragmatic trial demonstration projects: UH2 AT007797, UH3

DK102384 (TiME), UH2 MH106338-02, UH3 MH106338-02 (TSOS), UH3 AT007769 (ABATE), UH3 NS088731-02, UH3 AT007788-02 (PPACT), UH2 DK104655-02, UH3 DK104655-02 (ICD-Pieces), UH2 AT007766-01, UH3 AR066795 (LIRE), UH2 AG049619-02, UH3 AG049619-02 (PROVEN), UH3 AT007782-02, UH3 CA188640-02 (STOP CRC), and UH2 AT007755-01 (SPOT).

Disclosures: Disclosures can be viewed at www.acponline.org/authors/icmje/ConflictOfInterestForms.do?msNum=M17-0863.

Requests for Single Reprints: Gregory Simon, MD, MPH, Kaiser Permanente Washington Health Research Institute, 1730 Minor Avenue, #1600, Seattle, WA 98101; e-mail, simon.g@ghc.org. Current author addresses and author contributions are available at Annals.org. *Ann Intern Med.* doi:10.7326/M17-0863

References

1. Institute of Medicine. Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk. Washington, DC: National Academies Pr; 2015.
2. Warren E. Strengthening research through data sharing. N Engl J Med. 2016;375:401-3. [PMID: 27518656] doi:10.1056/NEJMp1607282
3. Krumholz HM, Terry SF, Waldstreicher J. Data acquisition, curation, and use for a continuously learning health system. JAMA. 2016; 316:1669-70. [PMID: 27668668] doi:10.1001/jama.2016.12537
4. Taichman DB, Backus J, Baethge C, et al. Sharing clinical trial data: a proposal from the International Committee of Medical Journal Editors [Editorial]. Ann Intern Med. 2016;164:505-6. [PMID: 26792258] doi:10.7326/M15-2928
5. Krumholz HM, Ross JS, Gross CP, et al. A historic moment for open science: the Yale University Open Data Access project and medtronic [Editorial]. Ann Intern Med. 2013;158:910-1. [PMID: 23778908] doi:10.7326/0003-4819-158-12-201306180-00009
6. Patel MR, Armstrong PW, Bhatt DL, et al; Academic Research Organization Consortium for Continuing Evaluation of Scientific Studies— Cardiovascular (ACCESS CV). Sharing data from cardiovascular clinical trials—a proposal. N Engl J Med. 2016;375:407-9. [PMID: 27518659] doi:10.1056/NEJMp1605260
7. Pencina MJ, Louzao DM, McCourt BJ, et al. Supporting open access to clinical trial data for researchers: the Duke Clinical Research Institute-Bristol-Myers Squibb Supporting Open Access to Researchers Initiative. Am Heart J. 2016;172:64-9. [PMID: 26856217] doi:10.1016/j.ahj.2015.11.002
8. OPTUM Labs. 2017. Accessed at www.optumlabs.com on 17 May 2017.

Current Author Addresses: Drs. Simon and Green: Kaiser Permanente Washington Health Research Institute, 1730 Minor Avenue, #1600, Seattle, WA 98101.
Drs. Coronado, DeBar, and Vollmer: Kaiser Permanente Center for Health Research, 3800 North Interstate Avenue, Portland, OR 97227.
Dr. Dember: Perelman School of Medicine, Blockley Hall, Office 920, 423 Guardian Drive, Philadelphia, PA 19104.

Dr. Huang: University of California, Irvine School of Medicine,101 The City Drive South, City Tower, Suite 400, Mail Code 4081, Orange, CA 92868.

Drs. Jarvik and Zatzick: Comparative Effectiveness, Cost and Outcomes Research Center, University of Washington, Box 359728, 325 Ninth Avenue, Seattle, WA 98104.

Dr. Mor: Brown University, Department of Community Health, Box G-S121-2, Providence, RI 02912.

Drs. Ramsberg and Platt: Harvard Pilgrim Health Care Institute and Harvard Medical School, Department of Population Medicine,401 Park Drive, Suite 401 East, Boston, MA 02215.

Dr. Septimus: Texas A&M College of Medicine, 6431 Fannin Street, MSE R.234, Houston, TX 77030. Ms. Staman: CHB Wordsmith, Inc., 22 Bagwell Avenue, Raleigh, NC 27607.

Dr. Vazquez: University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, TX 75390. Dr. Hernandez: Duke University School of Medicine, 3116 North Duke Street, Durham, NC 27704.

Author Contributions: Conception and design: G.E. Simon, L.L. DeBar, L.M. Dember, B.B. Green, V. Mor, E.J. Septimus, K.L. Staman, W.M. Vollmer, A.F. Hernandez, R. Platt. Analysis and interpretation of the data: G.E. Simon, L.L. De- Bar, L.M. Dember, B.B. Green, E.J. Septimus, M.A. Vazquez, A.F. Hernandez, R. Platt. Drafting of the article: G.E. Simon, L.L. DeBar, J. Ramsberg, E.J. Septimus, K.L. Staman, M.A. Vazquez, R. Platt.

Critical revision for important intellectual content: G.E. Simon, G. Coronado, L.L. DeBar, L.M. Dember, B.B. Green, S.S. Huang, J.G. Jarvik, V. Mor, E.J. Septimus, K.L. Staman, M.A. Vazquez, W.M. Vollmer, A.F. Hernandez, R. Platt. Final approval of the article: G.E. Simon, G. Coronado, L.L. DeBar, L.M. Dember, B.B. Green, S.S. Huang, J.G. Jarvik, V. Mor, J. Ramsberg, E.J. Septimus, K.L. Staman, M.A. Vazquez, W.M. Vollmer, D. Zatzick, A.F. Hernandez, R. Platt. Provision of study materials or patients: J.G. Jarvik, E.J. Septimus, D. Zatzick. Obtaining of funding: L.M. Dember, B.B. Green, J.G. Jarvik, A.F. Hernandez.