HHS Public Access

Author manuscript

Ann Intern Med. Author manuscript; available in PMC 2018 May 07.

Published in final edited form as:

Ann Intern Med. 2017 November 07; 167(9): 668–670. doi:10.7326/M17-0863.

Data Sharing and Embedded Research

Gregory E. Simon, MD, MPH,

Kaiser Permanente Washington Health Research Institute, Seattle, Washington

Gloria Coronado, PhD,

Kaiser Permanente Center for Health Research, Portland, OR

Lynn L. DeBar, PhD, MPH,

Kaiser Permanente Center for Health Research, Portland, Oregon

Laura M. Dember, MD,

Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA

Beverly B. Green, MD, MPH,

Kaiser Permanente Washington Health Research Institute, Seattle, Washington

Susan S. Huang, MD, MPH,

University of California Irvine School of Medicine, Orange, California

Jeffrey G. Jarvik, MD, MPH,

University of Washington, Seattle, WA

Vincent Mor, PhD,

Brown University, Providence, RI

Joakim Ramsberg, PhD,

Department of Population Medicine, Harvard Pilgrim Health Care Institute and Harvard Medical School

Edward J. Septimus, MD,

Research and Infectious Diseases Clinical Services Group, Hospital Corporation of America Nashville TN AND Texas A&M College of Medicine Houston TX

Karen L. Staman, MS,

CHB Wordsmith, Inc, Raleigh NC

Miguel A. Vazquez, MD,

University of Texas Southwestern Medical Center, Dallas, Texas

William M. Vollmer, PhD,

Kaiser Permanente Center for Health Research, Portland, Oregon

Address for correspondence: Gregory Simon MD MPH, Investigator, Kaiser Permanente Washington Health Research Institute Psychiatrist, Behavioral Health Service, Kaiser Permanente, Co-Chair, Scientific Advisory Board, Depression and Bipolar Support Alliance, 1730 Minor Ave. #1600, Seattle, WA 98101, Office: 206-287-2979, Fax: 206-287-2871, simon.g@ghc.org.

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

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

Douglas Zatzick, MD,

University of Washington School of Medicine, Seattle, Washington

Adrian F. Hernandez, MD, MHS, and

Duke University School of Medicine, Durham, North Carolina

Richard Platt, MD, MS

Department of Population Medicine, Harvard Pilgrim Health Care Institute and Harvard Medical School

Keywords

Data sharing; embedded research; pragmatic clinical trials

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). A number of solutions have been proposed to encourage the sharing of analyzable research datasets (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 (EHR), 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 still be substantial.

While 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 EHRs. 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 could differ from publicly reported quality measures.

In an ideal world of transparency regarding healthcare processes and outcomes, health systems would have no expectation of or need for privacy regarding quality of health care delivery. But the current world is not perfect, and unintentional disclosures from participation in embedded research could 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 of 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 individual 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 there is no analogous regulatory protection for providers, practices, and health systems participating in research there is a reasonable corollary. Such protections are especially important for providers included in cluster-randomized trials, where explicit provider consent is uncommon. The notion that health systems, providers, and/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 healthcare providers or 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 healthcare 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.

Data sharing solutions for embedded research

To motivate organizations to opt in to embedded research for the greater good, we must recognize that sharing patients' information can reveal sensitive information about providers or health systems. We recommend coupling that recognition with a framework for data sharing that champions making available as much of the data 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 re-identified, and reduce the risk either by modifying the data to be shared (e.g. 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:

- 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)

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

Balancing risks and public health value

More restrictive data sharing structures necessarily require greater resources. In comparison to a public archive, establishing a private archive requires personnel resources to review and approve users and specific uses. In comparison to a data archive, establishing a data enclave to respond to users' queries requires significantly 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 consideration of the following questions:

- What data could be shared by the least restrictive mechanism, a public archive open to any interested user?
- What additional data could be shared 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 NIH Health Care Systems Research Collaboratory were asked to consider these questions when creating a plan for sharing study data. Table 1 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.

Acknowledgments

Grant Support

This work was supported by a cooperative agreement [U54 AT007748] from the National Institutes of Health (NIH) Common Fund for the Coordinating Center of the NIH Health Care Systems Research Collaboratory and by the following grants from the NIH for the pragmatic trial demonstration projects: UH2 AT007797, UH3 DK102384 (TiME); UH2 MH106338-02; UH3 MH106338-02 (TSOS); UH3 AT007769 (ABATE); UH3 NS088731-02, UH3AT007788-02 (PPACT); UH2DK104655-02, UH3DK104655-02 (ICD-Pieces); UH2AT007766-01, UH3AR066795 (LIRE); UH2AG049619-02, UH3AG049619-02 (PROVEN); UH3AT007782-02, UH3CA188640-02 (STOP CRC); UH2AT007755-01 (SPOT).

References

1. Institute of Medicine (U.S.), Institute of Medicine (U.S.)., editor. Sharing clinical trial data: maximizing benefits, minimizing risk. Washington, D.C: National Academies Press; 2015. p. 290
2. Warren E. Strengthening Research through Data Sharing. New England Journal of Medicine. 2016 Aug 4; 375(5):401–3. [PubMed: 27518656]
3. Krumholz HM, Terry SF, Waldstreicher J. Data Acquisition, Curation, and Use for a Continuously Learning Health System. JAMA. 2016 Oct 25; 316(16):1669–70. [PubMed: 27668668]
4. Taichman DB, Backus J, Baethge C, Bauchner H, de Leeuw PW, Drazen JM, et al. Sharing Clinical Trial Data: A Proposal From the International Committee of Medical Journal Editors. Annals of Internal Medicine. 2016 Apr 5.164(7):505. [PubMed: 26792258]

5. Krumholz HM, Ross JS, Gross CP, Emanuel EJ, Hodshon B, Ritchie JD, et al. A historic moment for open science: the Yale University Open Data Access project and medtronic. *Ann Intern Med*. 2013 Jun 18; 158(12):910–1. [PubMed: 23778908]
6. The Academic Research Organization Consortium for Continuing Evaluation of Scientific Studies — Cardiovascular (ACCESS CV). Sharing Data from Cardiovascular Clinical Trials — A Proposal. *New England Journal of Medicine*. 2016 Aug 4; 375(5):407–9. [PubMed: 27518659]
7. Pencina MJ, Louzao DM, McCourt BJ, Adams MR, Tayyabkhan RH, Ronco P, 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. *American Heart Journal*. 2016 Feb;172:64–9. [PubMed: 26856217]
8. OPTUM Labs [Internet]. Website; Available from: <https://www.optumlabs.com> Accessed May 17, 2017

Table 1

Data Sharing Solutions Using Different Mechanisms*

Brief Trial Description	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 significant public health value?	Data Sharing Solution
PPACT Pain Program for Active Coping and Training (NCT02113592): Goal: 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 three health systems.	Naïve comparisons of opioid prescribing rates across providers or facilities.	Public dataset will not include facility or health system identifiers or patient-level variables likely to allow re-identification of providers or facilities.	No. The primary analysis can still be closely replicated without the additional data.	Public archive of a modified dataset
SPOT Suicide Prevention Outreach Trial (NCT02326883). Goal: Evaluate population-based outreach programs to prevent suicide attempt in high-risk outpatients. Will include ~16,000 patients in four health systems.	Naïve comparisons of suicide attempts or suicide mortality across health systems.	Public dataset will not include an indicator for health system.	Possibly. Datasets including health system identifiers will be available on request through a supervised data archive, subject to formal agreements regarding use and re-disclosure.	Public archive of a modified dataset
STOP CRC Strategies and Opportunities to Stop Colorectal Cancer in Priority Populations (NCT01742065) Goal: Evaluate systematic EHR embedded program for mailing of fecal immunochemical test kits to increase colorectal-cancer screening rates in Federally Qualified Health Centers (FQHCs). Will include ~41,000 patients in 26 FQHC clinics.	Naïve comparisons of cancer screening rates across providers or clinics. Perceived discrepancies between study outcome measures and publicly reported quality measures.	None.	Yes. De-identified patient-level data will be available on request, via a supervised data archive, subject to formal agreements regarding use, redisclosure, and data destruction.	Private archive managed by study team
ICD-Pieces Improving Chronic Disease management with Pieces (NCT02587936). Goal: 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 four health systems.	Naïve 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 National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) supervised data archive. Identifiers for healthcare system, primary practice and patients will be removed.	Private archive managed by NIDDK
ABATE Active Bathing to Eliminate Infection (NCT02063867) Goal: Evaluate the impact of daily antiseptic bathing, supplemented by nasal ointments for patients harboring methicillin-resistant <i>Staphylococcus aureus</i> (MRSA), on clinical isolates of multidrug-resistant organisms and bloodstream infections attributable to non-critical care hospital units. Will include over 600,000 patients in 53 hospitals.	Naïve comparison of infection or complication rates across facilities. Perceived discrepancies between study outcome measures and publicly reported quality measures. Disclosure of proprietary business information regarding utilization patterns.	None.	Yes. All individual-level data will remain behind the health system firewall. Following a private data enclave structure, potential users may propose specific queries, and only query results will be shared.	Private enclave managed by study team
LIRE Lumbar Image Reporting with Epidemiology (NCT02015455) Goal: Determine if inserting epidemiological benchmarks (essentially representing the	Naïve comparison of opioid prescribing rates and other treatments across	None.	Yes. Patient-level datasets will be deidentified by health systems, clinics,	Private archive managed by study team.

Brief Trial Description	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 significant public health value?	Data Sharing Solution
normal range) into lumbar spine imaging reports reduces subsequent tests and treatments. Will include four health systems, 100 clinics and ~245,000 patients.	primary care providers and health systems. Disclosure of proprietary business information regarding utilization patterns.		providers, and patients. Investigators will authorize release to specific users for specific purposes.	

* Assumes HIPAA-compliant patient de-identification for all patients and a data use agreement where appropriate.