Invited Forum: Challenges in Making Data Available
Tutorial

aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Practical Tips for Ethical Data Sharing

Advances in Methods and Practices in Psychological Science
2018, Vol. 1(1) 131–144

DOI: 10.1177/2515245917747656
www.psychologicalscience.org/AMPPS
SAGE

Michelle N. Meyer
Geisinger Health System, Danville, Pennsylvania

Abstract

This Tutorial provides practical dos and don'ts for sharing research data in ways that are effective, ethical, and compliant with the federal Common Rule. I first consider best practices for prospectively incorporating data-sharing plans into research, discussing what to say—and what not to say—in consent forms and institutional review board applications, tools for data de-identification and how to think about the risks of re-identification, and what to consider when selecting a data repository. Turning to data that have already been collected, I discuss the ethical and regulatory issues raised by sharing data when the consent form either was silent about data sharing or explicitly promised participants that the data would not be shared. Finally, I discuss ethical issues in sharing "public" data.

Keywords

morality, data sharing, IRB, research ethics, responsible conduct of research

Received 9/16/17; Revision accepted 11/21/17

In 2011, I attended the annual Social, Behavioral, and Educational Research Conference of Public Responsibility in Medicine and Research (PRIM&R). PRIM&R is essentially the guild for institutional review board (IRB) administrators and other research-oversight personnel and offers a Certified IRB Professional (CIP) credential along with best practices for IRB review of research involving human participants. That year, the conference organizers, during some introductory remarks, showed a slide with a quotation from an actual IRB submission: "After the study is completed," the slide read, "videotapes will be destroyed personally by the investigator with a sledgehammer." The exact purpose of that slide has been lost to memory, but presumably it was meant to rouse the early-morning audience with an amusing illustration of the lengths to which some exasperated researchers will go to assure their IRBs that participants' data will be protected.

Over the years, as I have watched the open-science movement blossom, that slide has come to illustrate, for me, something else: how far the IRB and research-ethics communities have to go in embracing data sharing. At the risk of stating the obvious, it is rather difficult to share data that have been sledgehammered to smithereens.

Why should researchers share their data? There are several legal, ethical, and practical reasons. Journals (e.g., Cozzarelli, 2004; *Nature*, 2017; *Science*, 2017), funders (e.g., National Institutes of Health, or NIH, Office of Extramural Research, 2007; National Science Foundation, or NSF, 2014, Article 44; PCORI, 2016), and professional societies (e.g., American Psychological Association, or APA, 2017, § 8.14) are increasingly requiring some form of data sharing. Even if a data-sharing clause is not explicitly included in a grant, researchers conducting publicly funded research arguably have an obligation to return the data they were paid to collect to the public realm. And even if research is not publicly funded, when a scientist publishes a claim about the world, he or she invites that claim to be tested by others through reanalysis and replication (Meyer & Chabris, 2014), activities that require access to the original data and methods, respectively. This obligation is even more critical in the wake of the "replication crisis," when the public's and funders' confidence in science appears to be fragile. Moreover, some scientific questions can be answered only with very large samples that require a consortium approach in which many researchers pool their data. Also, data sharing can be in researchers'

Corresponding Author:
Michelle N. Meyer, Geisinger Health System, Center for Translational Bioethics and Health Care Policy, 100 N. Academy Ave., Danville, PA 17822-2101
E-mail: mmeyer@geisinger.edu

self-interest, as there is some evidence that it leads to increased citation of the original research, at least in the case of clinical trials with cancer patients (Piwowar, Day, & Fridsma, 2007), gene-expression microarray studies (Piwowar & Vision, 2013), astronomy research (Henneken & Accomazzi, 2011), and astrophysics research (Drachen, Ellegaard, Larsen, & Dorch, 2016). And a demonstrable history of data sharing may be attractive to funders. Last—but not least—research participants are often motivated by their ability to contribute to science and want their data to be widely shared.

None of this is to say that, once one has decided to share data, the path forward is entirely straightforward. Any researcher who publishes should be prepared to immediately share data for the limited purpose of allowing other researchers to reproduce those published analyses. (Data should be shared publicly if at all possible, but may be shared only upon request if absolutely necessary to protect or keep promises to participants.) But reasonable people can disagree about when to share data for broader purposes, such as enabling other researchers to conduct new analyses or to combine the data with other data sets.[1] Data can be extraordinarily expensive and time-consuming to collect. And not every researcher is equally positioned to exploit a data set quickly before sharing; some have teams of graduate students and postdocs, whereas others work nearly entirely by themselves. Depending on the circumstances, it may be entirely acceptable for data collectors to embargo their data for a significant period of time, until they are able to produce one or more publications. (A probable exception is when the data are, say, medically actionable and withholding the data would directly harm people.) Reasonable people can also disagree about how secondary researchers should credit original data collectors.

In this Tutorial, I first offer several dos and don'ts for enabling newly collected data to be shared. I conclude with thoughts about what to do when one wants to share data that were previously collected without participants' explicit consent to data sharing.

Preparing to Share Data Effectively and Responsibly

DON'T promise to destroy your data

The strong default rule in science should be that research data will not be destroyed. Ordinarily, researchers should not volunteer to take a sledgehammer, or any other tool of destruction, to their data. And ordinarily, IRBs should not require the inclusion of data-destruction clauses in IRB applications, protocols, or consent forms. Neither the NIH nor NSF requires destruction of data, nor does the Common Rule (Federal Policy for the Protection of Human Subjects, 2017), the federal regulations that govern most federally funded research with human participants and strongly inform IRB review of even non-federally funded research.

There will, of course, be exceptions when data destruction is reasonable, but these should be rare, and any act or IRB requirement of data destruction should be explicitly justified. For instance, when participants' identities are no longer important for purposes of reproducing or replicating the research and the continued existence of the research data poses a very significant privacy risk to participants, then destroying identifiers (or the code linking identities to data) may be reasonable. Sometimes, raw data themselves are nearly inextricably linked to identity, as may be the case with the kind of video data that the nameless researcher mentioned in the opening paragraph pledged to smash. If participants were recorded, say, discussing illegal behavior, then destroying the video footage would likely be justified.

However, as I discuss later, there is a wide range of options for data sharing, from depositing data into a public repository open to all, to allowing access only by qualified researchers who have signed a strict data-use agreement. Even if researchers, for privacy reasons, never share their data with anyone else, retention can be important in allowing them to double-check the integrity of their original research and to defend their work if it is questioned (Neyfakh, 2015). In a world where safe-deposit boxes exist, raw data should be both highly identifiable and highly sensitive before the last resort of data destruction is contemplated.

DON'T promise not to share data

Too often, consent forms promise participants that their data "will be kept private and confidential to the extent permitted by law," or that "only the research team will have access" to the data. Such routine promises are often thoughtlessly included in modern consent forms that are adapted from earlier studies. Sometimes researchers may intentionally submit consent forms that promise the data will not be shared (or that are silent about data sharing) in an effort to obtain quicker IRB approval. This shortsighted strategy will cause considerable difficulties (which I discuss later) if the researcher later wishes to or (pursuant to evolving journal and funder requirements) must share data.

DON'T promise that research analyses of the collected data will be limited to certain topics

After promises to destroy data and promises not to share them, the next most problematic language found in many consent forms is language that suggests the

data will be used only for particular research purposes. Although the original researcher may never wish to conduct other analyses of the data, secondary researchers may well wish to do so. Original researchers should, to the extent possible, disclose how they themselves plan to use the data. But in asking participants to additionally consent to data sharing, original researchers should make it clear that other researchers may use the data for a variety of other purposes, up to and including any purpose at all, without recontacting participants or obtaining their consent to those new purposes.

DO get consent to retain and share data

Instead of promising to destroy or not to share data, researchers should build data-retention and data-sharing plans into IRB applications, experimental protocols, and consent forms. Researchers need not reinvent the wheel; several examples of data-sharing language (often approved by one or more IRBs) are available online and may be adapted as appropriate for different studies (see Databrary, n.d.-a, n.d.-b; Halchenko & Gorgolewski, 2015b, 2015c; Inter-university Consortium for Political and Social Research, 2017c; Murphy, 2016). Participants should be told what types of individuals will have access to their data: other researchers at the same institution, researchers at other institutions, commercial entities (and if so, whether participants will share in any resulting profits), governments, or the general public. They should also be told the purposes for which their data may be reused: for reanalysis and replication only or for new analyses (and if the latter, whether there will be any limits on the kinds of secondary analyses that may be conducted).

In making these disclosures, researchers should err on the side of obtaining participants' consent to broader and more public data sharing. If the data turn out to be more sensitive than anticipated, researchers retain flexibility to choose a more limited form of data sharing than the obtained consent permits. The converse, of course, is not true.

Tiered consent options can be used to provide participants with some control over how broadly their data are shared for secondary research purposes. The level of consent can vary along two different axes: That is, participants can be given a choice over whether to share some but not all of their data, and they can also be given a choice over whether to share their data with some groups but not all others. (Participants should generally *not* be given the option of withholding their data from other researchers who aim only to reproduce the original analysis, but should be told that their data may be shared for those purposes.) However, it will generally also be ethically acceptable if participants' only choice is to consent to their data being shared as described in the protocol (which may indicate very broad sharing) or not to participate in the study at all.

DO incorporate data-retention and -sharing clauses into IRB templates

Many IRBs have developed protocol and consent templates to help ensure that researchers address all critical aspects of their studies, as required by the Common Rule and institutional policy. Researchers may not be thinking about the eventuality of data sharing when their focus is on simply gaining approval to collect the data in the first place, but including data-sharing clauses in IRB templates would nudge researchers (and IRBs) toward data sharing and help reorient all parties from a culture of data secrecy to a culture of data sharing.

Templates are only defaults, and a data-sharing clause could be overridden when the IRB (or the researcher) believes that circumstances dictate doing so. But researchers and IRBs should not assume that data cannot ethically be retained and shared. Neither should they assume that individual participants or participant populations necessarily view their data as sensitive or—even if they do—believe that their data should be destroyed or kept secret by the primary research team. In general, it will be much more reasonable to ask questions about how and with whom data may be shared than to ask questions about whether it may be shared at all. Even highly sensitive, highly re-identifiable data, such as those collected through the Personal Genome Project, can be shared publicly if participants' comprehension of the risks is confirmed through brief quizzes administered during the consent process (Lunshof, Chadwick, Vorhaus, & Church, 2008). Consent comprehension quizzes can be used in other studies to ensure that participants understand the risks of a variety of levels of data sharing. With such safeguards in place, there should be no excuse for an IRB to prevent participants from making a knowing, voluntary decision to share their data.

DO be thoughtful when considering risks of re-identification

Two contrary impulses must both be avoided when data sharing is contemplated. First, it is natural for researchers to be enthusiastic about their research and—at least in the case of those who are laudably buoyed by the current open-science momentum—about sharing their data. But that eagerness, and the fact that re-identification is itself a specific domain of expertise, can prevent researchers from exercising necessary caution and reflection before sharing.

An "anonymous" data set, for instance, may easily cease to be anonymous if it includes variables that allow relatively unique individuals to be identified. A recent string of high-profile re-identification "attacks" by researchers has shown that it is possible to re-identify some data on the basis of, for example, full ZIP code, full birth date, and sex (Sweeney, 2002); Web search queries (Barbaro & Zeller, 2006); online movie reviews (Narayanan & Shmatikov, 2008); genomic data (Gymrek, McGuire, Golan, Halperin, & Erlich, 2013); cell-phone data (de Montjoye, Hidalgo, Verleysen, & Blondel, 2013); taxi-passenger data (Tockar, 2014); and credit-card metadata (de Montjoye, Radaelli, Singh, & Pentland, 2015).

Some data, although not easily re-identifiable by the public, are easily re-identifiable by people who know the participant. In some cases, that may be acceptable; in others, it may cause considerable harm. For instance, a hospital paid a $2.2 million fine for allowing a television crew to film and broadcast the treatment and subsequent death of an "unidentified" patient whose family recognized him during the broadcast (Ornstein, 2016). Similarly, some psychology research involves studying family members. If anonymized data are reported for pairs or other small groups, or via couple indicators, then one participant need only identify his or her own responses in order to identify those of another family member.[2]

On the other hand, it is important to avoid a second impulse, to overestimate the risk of re-identification. Re-identification attacks by researchers have received a great deal of media attention (some people would say media hype; Barth-Jones, 2012a, 2012b). Risk is the magnitude of harm discounted by the probability of that harm occurring, and a great deal of data collected under the auspices of psychological science could be re-identified without any significant harm being done to participants. The harm from re-identification of some kinds of data, such as health data, can be difficult to estimate to the extent that laws regarding discrimination and preexisting conditions are uncertain.

Estimating the probability of re-identification is difficult because it, too, is a moving target: As the amount of available data about an individual increases, any one data set about that individual becomes increasingly re-identifiable. More data about most of us is becoming available over time. Yet it is important to consider not only the technical feasibility of re-identification, which is where the bulk of attention has been placed, but also the incentives, or lack thereof, for people to seek to re-identify research data sets, as well as the costs to them of attempting to do so (Wan et al., 2015). To date, as far as we know, research data sets have been re-identified only by privacy researchers seeking to demonstrate the technical feasibility of doing so.

Notwithstanding this admonition not to overreact to re-identification risk, all reasonable measures should be taken to de-identify data except when the data are incontestably innocuous or participants have knowingly given clear consent to share identified or readily identifiable data. In the wake of the string of re-identification attacks I mentioned earlier, some critics have all but dismissed as worthless the de-identification tools outlined in the regulations implementing the Health Insurance Portability and Accountability Act (HIPAA; Standards for Privacy of Individually Identifiable Health Information, 2002, § 164.514(b)(2)), as well as other de-identification tools. Such criticism sweeps far too broadly. For instance, Sweeney's (2002) re-identification of Massachusetts Governor Bill Weld on the basis of his five-digit ZIP code, full date of birth, and sex occurred prior to, and indeed prompted revisions to, the safe-harbor provision of the HIPAA Privacy Rule (Standards for Privacy of Individually Identifiable Health Information, 2002). Prior to that revision, Sweeney offered a theoretical estimate that 87% of U.S. individuals could be re-identified on the basis of these three variables. She later testified, however, that if the same data set met HIPAA's safe-harbor provision—under which ZIP codes are limited to the first three digits and birth dates are limited to year of birth—only 0.04% of individuals could be re-identified (Barth-Jones, 2012a, 2012b). Similarly, a systematic review of known re-identification attacks on health data found that most "re-identified" data sets had not been properly de-identified according to current standards in the first place, weakening claims about the efficacy of re-identification techniques (El Emam, Jonker, Arbuckle, & Malin, 2011).

Researchers can also use a variety of anonymizing tools instead of or in addition to HIPAA's safe-harbor de-identification, which involves removing 18 identifiers. Other techniques include "masking" original data by replacing them with random data and "blurring" variables by sharing them at a reduced "resolution" (e.g., reporting age ranges instead of specific ages in years or larger geographic regions instead of ZIP codes). HIPAA's Privacy Rule itself permits a second approach to de-identification: expert determination, in which an appropriate expert uses "generally accepted statistical and scientific principles and methods" to render data not individually identifiable, so that the risk of re-identification is "very small" (Standards for Privacy of Individually Identifiable Health Information, 2002, §164.514(b)). However, most researchers—and most IRBs—lack the expertise to properly de-identify or obfuscate data by going beyond rote application of HIPAA's safe-harbor rules. As both the identifiability of data sets and the imperative to share data grow, the long-term solution may be to embed de-identification experts into research institutions, much as experts in statistics and survey methods now form standing "cores" that serve the research enterprise in many institutions.

In the short term, institutional privacy offices will tend to have more expertise in recognizing re-identification risks and in recommending solutions than will most IRBs. Helpful open-source de-identification tools also exist (Halchenko & Gorgolewski, 2015a; OpenfMRI, n.d.), and some data repositories review deposits for disclosure risks and offer de-identification and similar curation services (Inter-university Consortium for Political and Social Research, 2017a, 2017b).

DO consider working with a data repository

Researchers should strongly consider depositing their data in a repository rather than waiting to be asked for their data. In an effort to obtain data for reanalysis, Wicherts, Borsboom, Kats, and Molenaar (2006) e-mailed the corresponding authors of 141 articles published in APA journals. All authors who publish in these journals must sign the APA Certification of Compliance With APA Ethical Principles (APA, 2003), Principle 8.14 of which requires that psychologists share data with other "competent professionals who seek to verify the substantive claims through reanalysis." Wicherts et al. sent more than 400 e-mails, often including detailed descriptions of their study's aims, IRB approvals, signed assurances not to share the data further, and their curricula vitae. Yet after 6 months, 73% of the authors had still failed to share their data. Most of those authors explicitly refused or said they were unable to share, whereas others promised to share but did not or simply never responded to the requests. Only 11% of the authors shared their data after the first request.

Even if both data requestors and original data collectors are well intentioned, inertia by both parties may present an avoidable obstacle to efficient data sharing. Data repositories allow the original data collectors to provide maximum access by sharing once. Many repositories also enable preregistration, data analysis, posting of preprints, and sharing with lab members. They often provide other useful services as well, so that they offer one-stop shopping for the modern researcher.

DO be thoughtful when selecting a data repository

Researchers should consider the governance options available at different data repositories when selecting one, as a given repository may be more suitable for some data sets than for others (see Table 1). For instance, some repositories are entirely open, whereas others make data available only to "qualified researchers" (usually those who have registered an affiliation with a research institution, which may be asked to vouch for their research-ethics training and document that they have permission to conduct independent research). Limiting data access to qualified researchers excludes citizen scientists (and, at some institutions, trainees) and is controversial for that reason (The White House, 2016, p. 2). However, institutions can usually deter their affiliates from violating data-use agreements, whereas citizen scientists answer to no one, so restricted data sharing may be more appropriate for sensitive data; in those cases, less detailed versions of the same data sets may be made publicly available. Some repositories permit depositors to control the level of access to their data, and this control may include an option to make the data available to specific researchers via a private link. Also, some repositories have established data-use agreements or other terms of service that preclude, for instance, attempts to re-identify or recontact participants. Publications with sensitive data that are shared in a repository with documented processes for accessing such data are eligible for a special version of Open Science Framework's Open Data badge (Center for Open Science, n.d.).

Sharing Data That Were Previously Collected Without Explicit Consent to Share

So far, I have focused on best practices that, going forward, will bake data-sharing plans into IRB applications, protocols, and consent forms. But many researchers laudably wish (or are required) to share data that have already been collected via a consent form that either was silent about data sharing or promised that data would not be shared. What should researchers do in such cases?

Ethical considerations

Data sharing poses two risks to participants. One risk is that their data will be associated with their identity by someone they did not choose to share that identified data with; this can lead to harms, such as stigmatization and discrimination, in addition to basic loss of privacy. The other risk is that participants' data—even if not associated with their identity—will be used for research purposes to which they would not have consented, which would render them complicit in what they deem to be inappropriate research. The ethical and regulatory question is whether it is appropriate to impose these risks on participants, either without their explicit consent (when the consent form was silent about data sharing) or in contradiction to what they were promised when they gave their consent.

Whether data sharing in these circumstances is ethically appropriate or not must be determined on a

Table 1. Governance Attributes of Some Social-Science and General Data Repositories

Repository	Data type	Access tiers and licensing	Data-privacy constraints specified on the Web site	Data citation and other incentives
Databrary, https://nyu.databrary.org (New York University, with support from Pennsylvania State University)	Video, audio, and related metadata in the developmental and learning sciences	Five tiers are available: public, authorized users (data are available to users who are registered and have signed an access agreement cosigned by their home institution), excerpts (data are available to authorized users, who may show clips during presentations; see Gilmore, Kennedy, & Adolph, 2018, this issue), private (data are available only to collaborators), and unreleased (data are accessible only by the depositor). Use is limited to noncommercial research and educational purposes.	Sharing requires that participants signed the Databrary release template or an IRB has determined that the obtained consent contains equivalent language. In text documents, participants may not be identified by names or initials. IRB approval is required to upload personally identifiable data for sharing or to conduct research with personally identifiable data available on Databrary.	A Databrary-specific citation is generated for each volume when it is shared, and users agree to properly cite all Databrary resources used in their scholarly work.
Dryad, http://datadryad.org[a]	Content associated with scholarly research documents that are published, in press, or under review	Many journals integrate their submission process with Dryad. By default, data and other content associated with a scholarly research document are made public under a Creative Commons CC0 waiver upon online publication of that document. The availability of a postpublication embargo of data depends on journal policy. Some journals routinely allow postpublication embargoes of up to 1 year, in which case depositors may select this option at the time of submission. Dryad will facilitate longer embargoes, or embargoes of data associated with publications in other journals, with written permission from the journal editor or publisher.	Dryad specifies that "human subjects data must be properly anonymized and prepared under applicable legal and ethical guidelines."[b] Dryad permits no direct identifiers and no more than three indirect identifiers per data set. Although depositors are responsible for ensuring that these policies are met, every data package is reviewed by a curator for (among other things) the presence of personally identifiable or otherwise sensitive or inappropriate information.	Dryad assigns a DOI to each "data package" (all data files associated with a publication plus the metadata describing the set) and to individual data files.
figshare, http://figshare.com	Research data and other outputs (figures, theses, etc.) from any science field, in any file format, up to 5 GB; data deposited by an individual user should not be uploaded for commercial purposes, cannot have been previously published with a DOI, cannot have been copyrighted, and cannot contain broadly defined sensitive information	Data may be marked as private (accessible only to the uploader while logged in or to other people via a privately shared link) or public (under various Creative Commons licenses, some of which restrict use to noncommercial purposes or preclude alteration of the data file, figure, or other content). Additional Creative Commons and customizable licenses (including a restrictive-license template for sensitive data) are available through institutional figshare accounts.	Data uploaded by individual users may not contain "sensitive personal data," as defined by Section 2 of the U.K. Data Protection Act of 1998. Researchers depositing ethically sensitive data may choose to share only metadata.	figshare provides DOIs, including DOI reservations prior to publication. For each research output, figshare displays the number of views, downloads, and citations it has received, as well as its Altmetrics, and also enables in-browser visualization.

(*continued*)

Table 1. (*continued*)

Repository	Data type	Access tiers and licensing	Data-privacy constraints specified on the Web site	Data citation and other incentives
Harvard Dataverse, http://dataverse.harvard.edu (Institute for Quantitative Social Science, Harvard University)[c]	Quantitative and qualitative data in any format, from any discipline	The default license for all data sets is the Creative Commons waiver (CC0) that allows reuse of data without conditions. However, Dataverse offers other legally binding data-use and licensing agreements that depositors may require downloaders of their data to sign and also permits upload of customized agreements. Depositors may use a "Guestbook" feature to collect data about downloaders of open-access data. Although metadata are always open access, files themselves may be restricted use, in which case downloaders must be registered users.	Data may be uploaded only after the depositor has received all relevant approvals, including approval from an IRB, if required. Data uploaded by any one user must not contain information that would enable re-identification of any participants using the information available across that user's uploaded data sets and dataverses. Specifically, uploads cannot contain individuals' account numbers (e.g., Social Security numbers, credit-card numbers, medical-record numbers, health-plan numbers) or biometric identifiers (e.g., fingerprints, retina prints, voiceprints, DNA). Exceptions are allowed when identifiable information has already been made public, when the data are identifiable but reflect the public roles or other nonsensitive aspects of public figures, when a "sufficient length of time"[d] has passed since data collection, when participants have given explicit informed consent to public release of the data, and when the data pertain to deceased individuals (in the case of data created by a U.S. federal government agency or under a federal contract).	A citation that includes data-set version and a DOI is generated and automatically presented when a data set is created. In addition, tabular data uploaded in any of several standard formats are assigned a Universal Numerical Fingerprint and are automatically converted into files that can be downloaded in multiple formats. Tabular data can be analyzed within the Dataverse platform using the TwoRavens application, and data that include geospatial coordinates can be mapped using WorldMap.

(*continued*)

Table 1. (*continued*)

Repository	Data type	Access tiers and licensing	Data-privacy constraints specified on the Web site	Data citation and other incentives
ICPSR (Inter-university Consortium for Political and Social Research, University of Michigan), https://www.icpsr.umich.edu/icpsrweb[a]	Social and behavioral research data of all file types	The vast majority of ICPSR data holdings are public-use files with no restrictions on access beyond ICPSR's standard terms of use. However, in some cases, ICPSR provides vetted researchers and sponsor-supervised students access to restricted-use data versions that retain confidential or sensitive data. To request access to restricted-use data, an investigator must submit an application that describes the proposed research and includes a confidential data-security plan; a signed pledge of confidentiality; a restricted-data-use agreement signed by the user and a legal representative from his or her institution; and, in some cases, IRB approval or an exemption. Data may be provided in encrypted format, via a virtual data enclave, or (in the case of highly sensitive data) via a physical data enclave in Ann Arbor, MI.	Data containing identifying information may be deposited under conditions that are consistent with the consent form and IRB approval. Trained data curators review all deposited data to assess disclosure risk and, when necessary, either modify the data or restrict access to protect participants' confidentiality. Consultation on disclosure risk is available. Review of informed consent to ensure that data sharing is appropriate is also available.	To make data easier to use, ICPSR cleans and enhances data files and creates descriptive metadata. All data collections receive a data citation including a persistent DOI. When users cite the data (especially if they include the DOI), ICPSR is able to track the citations in a bibliography that allows other researchers to assess the impact and reuse of data collections.
OpenfMRI, http://openfmri.org (Stanford Center for Reproducible Neuroscience)	All forms of neuroimaging data that include MRI images and associated data	Unless otherwise noted, data are available under the Creative Commons CC0 1.0 license.	Data must be de-identified prior to sharing. Depositors are advised to receive either IRB approval to share the de-identified data or a determination from their IRB that sharing is allowed.	OpenfMRI provides a unique, permanent accession number for each data set. Users are encouraged to follow the Open Data Commons (n.d.) Attribution-Sharealike Community Norms and, in particular, to cite OpenfMRI in reports using data downloaded from OpenfMRI.

(*continued*)

Table 1. (*continued*)

Repository	Data type	Access tiers and licensing	Data-privacy constraints specified on the Web site	Data citation and other incentives
openICPSR (a self-publishing service operated by ICPSR),[e] https://www.openicpsr.org	Social and behavioral research data of all file types	Data at openICPSR are made available under an Attribution 4.0 Creative Commons license. Self-publishers choose to either make the data available for immediate public download or to restrict access. If access is restricted, users must apply for access and pay an administrative fee.	Self-publishers must attest that no individuals can be identified from information in the data collection. ICPSR does not conduct disclosure analysis on openICPSR collections.	Data collections are not curated by ICPSR and remain in their original format, although self-publishers may pay to have the data curated by ICPSR (i.e., cleaned and enhanced so they are easier to find and use). All openICPSR collections do receive a data citation including a persistent DOI.
OpenNeuro, https://openneuro.org (Stanford Center for Reproducible Neuroscience)	Neuroimaging data in Brain Imaging Data Structure (BIDS) format	Uploaded data are private (i.e., only collaborators can view and edit the data) for a limited time and then become public. Data and related analytic results are made publicly available under a Creative Commons CC0 or CC-BY license no later than 36 months following the first successful analysis of data from more than 1 participant; users may apply for up to two 6-month extensions.	Uploaded data must be owned by the depositor, who must have obtained necessary ethics permissions to share the data publicly. The data must not contain any HIPAA identifiers (e.g., names, ZIP codes, dates of birth, acquisition dates, facial features on structural scans).	OpenNeuro provides a unique, permanent accession number for each data set. It plans to add DOI functionality in the future. Data authors have free access to supercomputer analysis of their data via OpenNeuro's Web interface.
Open Science Framework, https://osf.io (Center for Open Science, Charlottesville, VA)	General science content, including data, materials, and code	Access may be public (depositors select from common licenses or upload their own) or private (accessible only to the depositor, contributors to the project or component, and users with a view-only link generated by the depositor).	Depositors should consult their institutions about HIPAA compliance concerns.	DOIs (no versioning) and Archival Resource Keys (ARKs) are available for public projects and registrations. Every project, component file, and user is assigned a unique, persistent URL that enables preformatted citations to be displayed on every Project Overview and Component Overview page. Each project can connect to preregistrations or preprints.

(*continued*)

Table 1. (*continued*)

Repository	Data type	Access tiers and licensing	Data-privacy constraints specified on the Web site	Data citation and other incentives
Zenodo, https://zenodo.org	Any research output (including multimedia) from any field; up to 50 GB per data set	Data may be marked as open, embargoed (data will become public at the end of a specified time), restricted (access is available only with the permission of the depositor), or closed. Depositors must specify a license for all publicly available files. Data are available for nonmilitary purposes only.	Data depositors are responsible for ensuring compliance with copyright, data-privacy, and other laws and for ensuring that the data are suitable for sharing.	Zenodo assigns DOIs to all publicly available uploads and also enables DOI versioning (i.e., a file can be modified after publication and users can cite different versions of a file or the entire set of files).

Note: This table summarizes information obtained through personal communication with representatives of the repositories in January 2018. For a comparative review of other aspects of data repositories, see Dataverse (2017). HIPAA = Health Insurance Portability and Accountability Act of 1996 (see Standards for Privacy of Individually Identifiable Health Information, 2002); IRB = institutional review board.

[a]Users pay a fee for this service, unless a journal or institution pays the fee. [b]See https://datadryad.org/pages/humanSubjectsData. [c]Dataverse, which was developed by Harvard University, is an open-source Web application for creating data repositories. Harvard Dataverse is a data repository that was made using that software and is open to data depositors and researchers worldwide. In the future, Harvard Dataverse plans to integrate with DataTags (https://datatags.org), which provides different transit, storage, and access options to support the sharing of data of varying degrees of sensitivity. The access and data-privacy policies described in this table reflect current options. [d]See Restriction 6 at https://dataverse.org/best-practices/harvard-dataverse-general-terms-use. [e]Curation services are not provided.

case-by-case basis. But in general, the argument for sharing will be stronger the more of the following conditions are met:

- The original consent form was merely silent about data sharing, and did not include a promise not to share data
- The data are not especially sensitive (i.e., re-identification would be unlikely to cause significant harm to participants)
- The data are not individually identified and are not especially likely to be re-identified (i.e., there are low incentives for anyone to re-identify the data or the data are unlikely to be re-identifiable alone or in combination with other available data sets)
- The shared data will be accessible only under restricted conditions, protected by agreements prohibiting re-identification
- Sharing will be limited to secondary research purposes that fall within the scope of the research described in the original consent form
- Sharing will be limited to secondary research purposes participants are not known to object to

Even when some of these considerations are not met, it is important to balance concerns about data privacy and data repurposing with the recognition that many participants prefer greater, rather than less, sharing of the data they contributed to science. Participants typically volunteer for research with the expectation that all reasonable efforts will be made to ensure that the results are correct, and data sharing for reanalysis and replication purposes helps to meet that objective. Also, participants who are members of groups that traditionally have been underrepresented in research may have a particular interest in having their data used widely (although their data may, for similar reasons, be more vulnerable to re-identification than other participants' data are). An especially strong case exists for nonconsensual data sharing for the limited purpose of reanalysis. In approving original research, IRBs must determine that the risks to participants are reasonable relative to the expected benefits of the research (Federal Policy for the Protection of Human Subjects, 2017, § 46.111(a)(2)). Those expected benefits may include direct benefits to participants, but given the IRB system's view of what constitutes a research-related benefit (e.g., incentives such as gift cards do not count; Meyer, 2013, pp. 276–279), the benefits of psychological research are likely to take the form of knowledge that is reasonably expected to result. Research analyses that cannot be reproduced because data cannot be shared arguably fail to qualify as knowledge at all, much less valuable knowledge. Similarly, it is a tenet of research ethics that research that is not well designed to rigorously answer an important question is unethical, because it means that any research-related risk (even, some people would say, the modest burden of time spent by participants) is necessarily wasted (Emanuel, Wendler, & Grady, 2000). Today, it is clear that scientific rigor and integrity require routine reanalysis and replication, which in turn require data sharing for at least those purposes.

Regulatory considerations

Except for data that are subject to HIPAA, data sharing exists in a sort of regulatory twilight zone. The Common Rule does not prohibit data sharing and is—or should be—no obstacle to consensual data sharing. Moreover, under the Common Rule, secondary research using shared data that are neither identified nor "identifiable"—that is, data from individuals whose identity cannot be "readily ascertained" (Federal Policy for the Protection of Human Subjects, 2017, § 46.102(e)), either directly or indirectly, through coding systems (Office for Human Research Protections, 2008)—does not constitute human-participants research. (Note that this narrow regulatory definition of "identifiable" ignores other methods of re-identification.) As a result, one prominent advisory body has concluded that it is not a Common Rule violation for an investigator to conduct secondary research on nonidentifiable data when that research falls outside the scope of the original obtained consent (Secretary's Advisory Committee on Human Research Protections, 2011, III, FAQ #3).

But what about the act of data sharing itself? Data sharing alone does not constitute human-participants research, and most retrospective data sharing will occur after a research protocol is closed out by an IRB, assuming that the original research was not exempt from IRB review in the first place (Federal Policy for the Protection of Human Subjects, 2017, § 46.104). But there is something artificial about separating the act of data sharing from the rest of a research study's trajectory, even if data sharing is contemplated only after the fact. IRBs review preresearch recruitment plans, so there is no particular reason why they could not review postresearch data-sharing plans (leaving aside the important fact that most IRBs are far less qualified to review data-sharing plans than they are to review recruitment plans). Certainly, institutions can implement policies that empower their IRBs to review data-sharing plans, even if data sharing is not covered by the Common Rule. Moreover, sharing data that were collected using a consent form that promised the data would not be shared likely constitutes a protocol violation. Researchers should therefore always consult their IRBs

before sharing data when participants were promised otherwise. If the incremental risk of data sharing above and beyond the risks to which participants already consented is minimal, and if certain protections are in place, an IRB may approve an amendment to the protocol to allow data sharing without recontacting participants and obtaining their consent for the new purpose (which is often infeasible).

Sharing "Public" Data

One final comment regarding sharing data with repositories is in order. The Common Rule does not consider nonintervention research to involve human participants unless the data obtained are not only identifiable but also "private"—that is, data "about behavior that occurs in a context in which an individual can reasonably expect that no observation or recording is taking place" or data that have "been provided for specific purposes by an individual and that the individual can reasonably expect will not be made public (e.g., a medical record)" (Federal Policy for the Protection of Human Subjects, 2017, § 46.102(e)(4)).

Expectations of privacy for tweets and public Facebook posts are evolving as media routinely republish or broadcast this content (sometimes with identities intact, sometimes with identities blurred). But existing data found on unlocked Twitter accounts and on Facebook posts set to "public" surely fail to meet the Common Rule's definition of "private." As a result, neither analyzing those data nor resharing them by depositing them in a public repository constitutes human-participants research subject to IRB review under the Common Rule. Nevertheless, aggregating otherwise disparate bits of public data in one analyzable data set amplifies attention to the information that users disclosed and enables inferences about individuals that they may not have predicted or intended. It also creates a permanent record that will persist even if those individuals delete their original posts. Researchers collecting sensitive public data should therefore consider whether it is appropriate to de-identify those data, especially if identities are not critical to them.

More troubling is the possibility that some researchers consider to be public data that they are able to access only by using false pretenses to join a closed community in which the data are shared for specific purposes. In 2016, for instance, researchers scraped data from more than 68,000 user profiles on the dating site OkCupid.com. The data set included username, age, sex, gender, sexual orientation, and location. It also included users' answers to 2,543 questions probing their political, religious, and moral beliefs; masturbatory habits; risk-taking (including illegal) behaviors; and sexual preferences. The researchers used responses to 14 of these questions to infer users' general cognitive ability and uploaded the data to a repository where it was available to anyone. When asked, the lead researcher responded that they had made no attempt to de-identify the data set, citing the fact that it was "already public" (Hackett, 2016, comment by E. Kirkegaard). (After ethical questions were raised about the data set, the repository first password-protected the files and then, following OkCupid's notice of copyright violation, removed them entirely.)

At the time, portions of OkCupid user profiles, including information on age, gender, and sexual orientation, were indeed publicly accessible through standard search-engine queries (that no longer appears to be the case). But answers to the survey questions were accessible only to people who had created an OkCupid account and answered the same questions. Users admittedly could set certain survey answers to "private," in which case they were accessible only to the company for use in its matching algorithm. But the fact that users were willing to disclose personal information to fellow members of a particular community, for a particular purpose (finding appropriate matches and being transparent with potential dates about their preferences), does not mean that they would have agreed to share the same information with researchers, much less with the public, and much less in a permanent data repository. The researchers appear to have been able to access those sensitive, re-identifiable data only by signing up for an OkCupid account under the pretense that they shared the purpose that brought that community together.

Conclusion

Psychological science has borne the brunt of negative publicity concerning the replication crisis. But it is also leading the way toward more rigorous, reproducible science. One important tool in the reproducibility tool kit is data sharing, which enables reanalysis, replication, and well-powered consortium science. Historically, IRBs and many researchers have prioritized data secrecy over data sharing. Participants do often have privacy interests that are important to consider. Consequently, they should be asked for their permission to share their data, and care should be taken in deciding how and with whom their data are shared. But it is past time for the research community to realize that participants typically also expect that the data they contribute will be used to advance scientific truth, not merely to make scientific claims that cannot be verified.

Action Editor

Daniel J. Simons served as action editor for this article.

Author Contributions

M. N. Meyer is the sole author of this article and is responsible for its content.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Notes

1. For instance, the Open Science Framework (OSF) awards its Open Data badge to researchers who make their data "publicly available on an open-access repository," but only those data that are "needed to reproduce the reported results" must be included (OSF, 2016, Criteria 1 and 2).
2. I thank reviewer Paul W. Eastwick for this example.

References

American Psychological Association. (2003). *Certification of compliance with APA ethical principles*. Retrieved from https://www.apa.org/pubs/authors/ethics02.pdf

American Psychological Association. (2017). *Ethical principles of psychologists and code of conduct*. Retrieved from http://www.apa.org/ethics/code/

Barbaro, M., & Zeller, T. (2006, August 9). A face is exposed for AOL searcher No. 4417749. *The New York Times*. Retrieved from http://www.nytimes.com/2006/08/09/technology/09aol.html?mcubz=0

Barth-Jones, D. (2012a, August 10). The debate over 're-identification' of health information: What do we risk? [Web log post]. Retrieved from http://healthaffairs.org/blog/2012/08/10/the-debate-over-re-identification-of-health-information-what-do-we-risk/

Barth-Jones, D. (2012b). *The 're-identification' of Governor William Weld's medical information: A critical reexamination of health data identification risks and privacy protections, then and now*. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2076397

Center for Open Science. (n.d.). *8. Approved protected access repositories*. Retrieved from https://osf.io/tvyxz/wiki/8.%20Approved%20Protected%20Access%20Repositories/

Cozzarelli, N. R. (2004). UPSIDE: Uniform principle for sharing integral data and materials expeditiously. *Proceedings of the National Academy of Sciences, USA*, *101*, 3721–3722.

Databrary. (n.d.-a). *Sample participant release script*. Retrieved from https://www.databrary.org/access/guide/investigators/release/asking/script.html

Databrary. (n.d.-b). *Video data release template, participants*. Retrieved from https://www.databrary.org/access/policies/release-template.html

Data Protection Act, c. 29 Part 1 §2 (1998). Retrieved from https://www.legislation.gov.uk/ukpga/1998/29/section/2

Dataverse. (2017). *A comparative review of various data repositories*. Retrieved from https://dataverse.org/blog/comparative-review-various-data-repositories

de Montjoye, Y.-A., Hidalgo, C. A., Verleysen, M., & Blondel, V. D. (2013). Unique in the crowd: The privacy bounds of human mobility. *Scientific Reports*, *3*, Article 1376. doi:10.1038/srep01376

de Montjoye, Y.-A., Radaelli, L., Singh, V. K., & Pentland, A. "S." (2015). Unique in the shopping mall: On the reidentifiability of credit card metadata. *Science*, *347*, 536–539.

Drachen, T. M., Ellegaard, O., Larsen, A. V., & Dorch, S. B. F. (2016). Sharing data increases citations. *LIBER Quarterly*, *26*, 67–82.

El Emam, K., Jonker, E., Arbuckle, L., & Malin, B. (2011). A systematic review of re-identification attacks on health data. *PLOS ONE*, *6*(12), Article e28071. doi:10.1371/journal.pone.0028071

Emanuel, E. J., Wendler, D., & Grady, C. (2000). What makes clinical research ethical? *Journal of the American Medical Association*, *283*, 2701–2711.

Federal Policy for the Protection of Human Subjects, 45 C.F.R. § 46 (2017).

Gilmore, R. O., Kennedy, J. L., & Adolph, K. E. (2018). Practical solutions for sharing data and materials from psychological research. *Advances in Methods and Practices in Psychological Science*, *1*, 121–130.

Gymrek, M., McGuire, A. L., Golan, D., Halperin, E., & Erlich, Y. (2013). Identifying personal genomes by surname inference. *Science*, *339*, 321–324.

Hackett, R. (2016, May 18). Researchers caused an uproar by publishing data from 70,000 OkCupid users. *Fortune*. Retrieved from http://fortune.com/2016/05/18/okcupid-data-research/

Halchenko, Y., & Gorgolewski, C. F. (2015a). *Anonymization tools*. Retrieved from https://open-brain-consent.readthedocs.io/en/latest/anon_tools.html

Halchenko, Y., & Gorgolewski, C. F. (2015b). *Sample consent forms*. Retrieved from https://open-brain-consent.readthedocs.io/en/latest/samples.html#chap-consent-samples

Halchenko, Y., & Gorgolewski, C. F. (2015c). *Ultimate consent form*. Retrieved from https://open-brain-consent.readthedocs.io/en/latest/ultimate.html#chap-consent-ultimate

Henneken, E. A., & Accomazzi, A. (2011). Linking to data—effect on citation rates in astronomy [Abstract]. In P. Ballester (Ed.), *Proceedings of ADASS XXI. Book Series: Astronomical Society of the Pacific Conference Series 461* (pp. 763–766). Retrieved from http://arxiv.org/abs/*1111*.3618

Inter-university Consortium for Political and Social Research. (2017a). *Confidentiality*. Retrieved from http://www.icpsr.umich.edu/icpsrweb/content/datamanagement/confidentiality/index.html

Inter-university Consortium for Political and Social Research. (2017b). *Guide to social science data preparation and archiving. Phase 5: Preparing data for sharing*. Retrieved from https://www.icpsr.umich.edu/icpsrweb/content/deposit/guide/chapter5.html

Inter-university Consortium for Political and Social Research. (2017c). *Recommended informed consent language for data sharing*. Retrieved from https://www.icpsr.umich.edu/icpsrweb/content/datamanagement/confidentiality/conf-language.html

Lunshof, J. E., Chadwick, R., Vorhaus, D. B., & Church, G. M. (2008). From genetic privacy to open consent. *Nature Reviews Genetics*, *9*, 406–411.

Meyer, M. N. (2013). Regulating the production of knowledge: Research risk–benefit analysis and the heterogeneity problem. *Administrative Law Review*, *65*, 237–298.

Meyer, M. N., & Chabris, C. (2014, July 31). Why psychologists' food fight matters. *Slate*. Retrieved from http://www.slate.com/articles/health_and_science/science/2014/07/replication_controversy_in_psychology_bullying_file_drawer_effect_blog_posts.html

Murphy, S. C. (2016). *Open science ethics templates*. Retrieved from https://github.com/seanchrismurphy/Open-science-ethics-templates

Narayanan, A., & Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. In *Proceedings of the 2008 IEEE Symposium on Security and Privacy* (pp. 111–125). Los Alamitos, CA: Institute of Electrical and Electronics Engineers.

National Institutes of Health, Office of Extramural Research. (2007). *NIH data sharing policy*. Retrieved from https://grants.nih.gov/grants/policy/data_sharing/

National Science Foundation. (2014). *Grant General Conditions (GC-1)*. Retrieved from https://www.nsf.gov/pubs/policydocs/gc1/feb14.pdf

Nature. (2017). *Availability of data, material and methods*. Retrieved from http://www.nature.com/authors/policies/availability.html#data

Neyfakh, L. (2015, June 18). The ethics of ethnography. *Slate*. Retrieved from http://www.slate.com/articles/news_and_politics/crime/2015/06/alice_goffman_s_on_the_run_is_the_sociologist_to_blame_for_the_inconsistencies.html

Office for Human Research Protections. (2008). *Coded private information or specimens use in research, guidance (2008)*. Retrieved from https://www.hhs.gov/ohrp/regulations-and-policy/guidance/research-involving-coded-private-information/index.html

Open Data Commons. (n.d.). *ODC Attribution-Sharealike Community Norms*. Retrieved from https://opendatacommons.org/norms/odc-by-sa/

OpenfMRI. (n.d.). *Information on de-identification of fMRI data*. Retrieved from https://openfmri.org/de-identification/

Open Science Framework. (2016). *1. view the badges*. Retrieved from https://osf.io/tvyxz/wiki/1.%20View%20the%20Badges/

Ornstein, C. (2016, April 21). New York hospital to pay $2.2 million over unauthorized filming of 2 patients. *The New York Times*. Retrieved from https://www.nytimes.com/2016/04/22/nyregion/new-york-hospital-to-pay-fine-over-unauthorized-filming-of-2-patients.html

PCORI. (2016). *Policy for data access and data sharing: Draft for public comment*. Retrieved from https://www.pcori.org/sites/default/files/PCORI-Data-Access-Data-Sharing-DRAFT-for-Public-Comment-October-2016.pdf

Piwowar, H. A., Day, R. S., & Fridsma, D. B. (2007). Sharing detailed research data is associated with increased citation rate. *PLOS ONE*, *2*(3), Article e308. doi:10.1371/journal.pone.0000308

Piwowar, H. A., & Vision, T. J. (2013). Data reuse and the open data citation advantage. *PeerJ, 1*, Article e175. doi:10.7717/peerj.175

Science. (2017). *Editorial policies*. Retrieved from http://www.sciencemag.org/authors/science-editorial-policies

Secretary's Advisory Committee on Human Research Protections. (2011). *Attachment D: FAQ's terms and recommendations on informed consent and research use of biospecimens*. Retrieved from https://www.hhs.gov/ohrp/sachrp-committee/recommendations/2011-october-13-letter-attachment-d/

Standards for Privacy of Individually Identifiable Health Information, 45 C.F.R. § 160 and 164 (2002).

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems*, *10*, 557–570.

Tockar, A. (2014, September 15). *Riding with the stars: Passenger privacy in the NYC taxicab dataset*. Retrieved from https://research.neustar.biz/2014/09/15/riding-with-the-stars-passenger-privacy-in-the-nyc-taxicab-dataset/

Wan, Z., Vorobeychik, Y., Xia, W., Clayton, E. W., Kantarcioglu, M., Ganta, R., . . . Malin, B. A. (2015). A game theoretic framework for analyzing re-identification risk. *PLOS ONE*, *10*(3), Article e0120592. doi:10.1371/journal.pone.0120592

The White House. (2016). *Precision medicine initiative: Data security policy principles and framework*. Retrieved from https://allofus.nih.gov/sites/default/files/PMI_Security_Principles_and_Framework_FINAL_022516.pdf

Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist*, *61*, 726–728.