Exchanging words: Engaging the challenges of sharing qualitative research data

James M DuBois  
*Washington University School of Medicine in St. Louis*

Jessica Mozersky  
*Washington University School of Medicine in St. Louis*

Meredith Parsons  
*Washington University School of Medicine in St. Louis*

Heidi A Walsh  
*Washington University School of Medicine in St. Louis*

Annie Friedrich  
*Washington University School of Medicine in St. Louis*

See next page for additional authors

Follow this and additional works at: [https://digitalcommons.wustl.edu/oa_4](https://digitalcommons.wustl.edu/oa_4)

Part of the Medicine and Health Sciences Commons

Please let us know how this document benefits you.

**Recommended Citation**

DuBois, James M; Mozersky, Jessica; Parsons, Meredith; Walsh, Heidi A; Friedrich, Annie; and Pienta, Amy, "Exchanging words: Engaging the challenges of sharing qualitative research data." Proceedings of the National Academy of Sciences of the United States of America. 120, 43. e2206981120 (2023).  
https://digitalcommons.wustl.edu/oa_4/3233

This Open Access Publication is brought to you for free and open access by the Open Access Publications at Digital Commons@Becker. It has been accepted for inclusion in 2020-Current year OA Pubs by an authorized administrator of Digital Commons@Becker. For more information, please contact vanam@wustl.edu.
In January 2023, a new NIH policy on data sharing went into effect. The policy applies to both quantitative and qualitative research (QR) data such as data from interviews or focus groups. QR data are often sensitive and difficult to deidentify, and thus have rarely been shared in the United States. Over the past 5 y, our research team has engaged stakeholders on QR data sharing, developed software to support data deidentification, produced guidance, and collaborated with the ICPSR data repository to pilot the deposit of 30 QR datasets. In this perspective article, we share important lessons learned by addressing eight clusters of questions on issues such as where, when, and what to share; how to deidentify data and support high-quality secondary use; budgeting for data sharing; and the permissions needed to share data. We also offer a brief assessment of the state of preparedness of data repositories, QR journals, and QR textbooks to support data sharing. While QR data sharing could yield important benefits to the research community, we quickly need to develop enforceable standards, expertise, and resources to support responsible QR data sharing. Absent these resources, we risk violating participant confidentiality and wasting a significant amount of time and funding on data that are not useful for either secondary use or data transparency and verification.

The NIH has published a policy enhancing already existing requirements to share data from NIH-funded research (1). The policy, which went into effect in January 2023, applies to both quantitative and qualitative research (QR) data (2). QR methods—such as interviews or focus groups—are frequently used to gather sensitive data (3–5). Until now, QR data have rarely been shared in the United States (6). While NIH may be leading the way in the United States, we believe that this is the way of the future: In August 2022, the White House Office of Science and Technology Policy issued a memorandum to all federal agencies instructing them to update their data-sharing policies to require sharing of all federally funded research data without embargos or costs to users by 2025 (7). The NSF has invested heavily in establishing the first US data repository specifically for qualitative data—the Qualitative Data Repository (QDR) at the University of Syracuse (NSF2116935).

In 2017, we published in Qualitative Psychology an article that asked, “Is it time to share QR data?” (8). We noted that researchers in several nations—e.g., Australia, Finland, and the United Kingdom—commonly share QR data. Doing so offers the same benefits as sharing other kinds of research data—enabling secondary analyses and synthesis reviews, increasing the impact of data collection in a cost-effective manner, providing trainees with real-world datasets to use when learning analytic skills, and fostering public trust in science through increased transparency (8, 9). Most qualitative researchers would balk at the idea of “reproducibility” given the inherent subjectivity or “researcher degrees of freedom” found in QR (4, 10). But transparency with raw data (e.g., interview transcripts) and codebooks is feasible, supports trust, and enables others to verify the warrant for claims made about the data (8, 9, 11, 12).

Our article was published along with three invited commentary articles. One set of commentators praised the idea, noting that QR data sharing would “raise the quality of qualitative methods” (13), while the other two commentaries strenuously objected, claiming a risk of “blackmail or murder” of research participants (14) and that repositories might become tools “of exploitation and scientific racism” if they support research without accountability to participant communities (15). We conceded that in specific contexts, these risks are plausible and not all QR data should be shared; but we do not believe that is the norm (16). As we note below, our ongoing research on QR data sharing has confirmed that these polarized views are characteristic of qualitative researchers in the US.

Later in 2017, we received funding from the NIH for a project, “Sharing Qualitative Research Data: Identifying and Addressing Ethical and Practical Barriers” (R01HG009351). We anticipated that already existing data-sharing requirements could—at any time—be enforced regarding QR data and felt that we were utterly unprepared in the United States (17). From an implementation science perspective, the research community in the United States had not identified
barriers and facilitators of QR data sharing, had not engaged stakeholders, did not have key resources for data deidentification, and did not have guidelines (including answers to important regulatory questions). Thus, we proposed several activities, which we recently completed:

- to interview 120 stakeholders (30 each of institutional review board (IRB) members, QR participants, data repository staff, and qualitative researchers),
- to conduct a broad national survey of qualitative researchers,
- to develop guidelines for QDS,
- to produce software that would assist with data deidentification.

The project has culminated in piloting data deidentification using our software and depositing 30 QR datasets with the largest social science data repository in the United States, the ICPSR at the University of Michigan.

While our main findings have been previously published, it is worth revisiting a few key findings. From our large stakeholder engagement interview study (N = 120), we found the following. IRB members recruited from research-intensive universities throughout the United States were generally supportive unless consent forms prohibited data sharing or data were not deidentified (18). QR participants recruited through our university’s research participant registry were generally supportive as long as data were deidentified and shared only with other researchers (19). Data repository staff, who were recruited internationally through the Open Access Directory, were unequivocally supportive but many felt unprepared to support researchers (18). Qualitative researchers who were recruited nationally and represented diverse fields were mostly concerned or ambivalent—often worrying about what IRBs or participants would think—while admitting they knew little about how it would work (18). Data from this project have been deposited with the ICPSR (20).

Our survey (N = 425) found that qualitative researchers in the United States are nearly evenly divided into those who oppose (41%) and those who support (49%) sharing QR data, with health science researchers (n = 152) generally more supportive than researchers from anthropology and sociology (n = 133) and other fields (n = 118). However, concerns about permission, participant trust, data sensitivity, IRBs, and data deidentification were shared by nearly all (79% or more). Crucially, 96% had never shared QR data with a repository (21).

We developed the Qualitative Data Sharing (QuaDS) software, which flags all Health Insurance Portability and Accountability Act (HIPAA) safe harbor identifiers as well as a series of potentially identifying variables (PIVs) such as institutions, race and ethnicity, rare diseases, and sexual orientation/LGBTQI+ status with excellent precision and recall (F = 0.96) (22). See Box 1 for further details.

Finally, 28 pilot project participants have deidentified and deposited data with the ICPSR. Funding sources for the original studies were diverse; most studies focused on health, sexual behavior, drug use, or involvement in the legal system. Links to all project materials and extensive guidance on QR data sharing can be found at QDStoolkit.org.

Drawing upon data from our project and the experience of working with qualitative researchers from diverse institutions as they prepared data for deposit, we address some common and important practical questions regarding qualitative data sharing. In what follows, we are guided by the assumption that data deposits must be responsible (e.g., not harm participants) and useful (e.g., support high-quality secondary use or data verification).

Karcher and colleagues define “epistemically responsible reuse (ERR)” as secondary use of qualitative data that “actively aims to understand as much of the original data and context as possible and does not make claims beyond what can be justifiably inferred from the data” (23, p. 1999). This is an appropriate goal for the collaborative work of repositories and researchers who deposit data. In addition, FAIR principles for all data sharing support the usefulness of shared data: Findability, Accessibility, Interoperability, and Reusability (24). These principles were developed in 2016

Box 1. Description of the QuaDS Software

QuaDS Software searches text files and ...

flags instances of HIPAA Safe Harbor Identifiers, including:

- Names
- Addresses
- Phone numbers
- Email addresses
- Birth dates

and instances of PIVs that may be used to identify someone inferentially when combined with other information:

- Race and ethnicity
- LGBTQI+ identification
- Institution names
- Geographical locations
- Rare diseases
- Numbers larger than 1*

QuaDS software performed with excellent precision/specificity (0.95) and recall/sensitivity (0.96) in a dataset consisting of 286,340 words from 70 qualitative interviews and personal narratives (22).

Flagged text is color coded to indicate whether the text likely requires replacement or is likely ok to leave.

Users can ignore flagged text or replace all or single instances. QuaDS generates a change log.

Further information on QuaDS can be found at qdstoolkit.org.

*We encourage review of numbers to screen for potentially identifying outlier values, e.g., "I was one of twelve children" or "I weighed 430 lb."
with the purpose of “enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals” (24). We will explore how these principles apply to QR data sharing as we examine key questions.

**Practical Questions about Qualitative Data Sharing**

Our data showed that nearly all qualitative researchers lack experience sharing data with a data repository. That means that they will have many questions as they attempt this complex activity, one that is fraught with scientific, logistical, regulatory, and ethical issues. Here, we briefly explore eight questions that must be answered when sharing data.

We focus on two of the most common forms of QR data collection—interviews and focus groups, which typically generate audio or audio–visual recordings and text transcripts (3). Other forms of data collection, e.g., through video recordings or ethnographic observation and field notes, are acknowledged as presenting special challenges for QR data sharing (4). For example, field notes often include reflexive notes by researchers on how they feel about a situation, which can be highly personal and might reflect poorly on individuals observed, and within a network of thick relationships, deidentification is impossible. Of course, similar challenges are faced when simply publishing findings in ethnography, and ethnography has a long history of sharing data and artifacts from projects, as observed in the “salvage ethnography” of Franz Boas (25). Similarly, video recordings pose an intrinsic challenge as they include at least two HIPAA safe harbor identifiers—voice prints and facial images—that are essential to the medium. Even so, in some cases, video data may pose minimal risks and participants may agree to sharing. The Databrary repository provides “a video data library for social scientists,” demonstrating that such sharing can be done (26). Nevertheless, we focus on transcripts as the most ubiquitous form of qualitative data and “low-hanging fruit” for data sharing.

**How Will QR Data Sharing Affect Project Planning?** Sharing data in accord with ERR and FAIR requires a substantial investment of time and planning (27). We focus here on three elements of planning.

**Budgeting and milestones.** Data sharing should be treated like any other project milestone when developing budget and staffing plans. Federal funding agencies allow investigators to budget costs related to data sharing. This may include costs related to “curating data, developing supporting documentation, formatting data …, deidentifying data, preparing metadata to foster discoverability, interpretation and reuse, …[and] data deposit fees” (28). However, all costs must be incurred during the project period (28). Thus, if someone waits to share data until a project is complete and closed, they risk bearing the costs of data sharing.

**Consent forms.** It is becoming commonplace for consent forms to address data sharing. However, QR researchers may be inclined to change template language. In the past, it was common to offer reassurances regarding confidentiality with statements such as “once data are analyzed they will be destroyed” or “no one outside of the research team will see your data” (29). These statements are generally not compatible with sharing data with a repository. Although some institutions might still permit sharing data if they are deidentified (because they no longer consider such data to be human subjects data), such statements do not serve the goal of informing participants in a transparent manner. Participants should be explicitly informed that data will be shared (29). Although saying more may restrict what can be done with data—and many institutions are loath to permit changes to template language—we believe it is best to provide additional details from the specific data-sharing plan. For example, it is appropriate to disclose whether only deidentified data will be shared, where data will be deposited, whether access to data will be controlled or restricted, and whether there are limitations on secondary uses. This approach is consistent with the sample consent information NIH has produced (30).

**Data management and documentation.** Data sharing is made significantly easier when best practices for data organization, storage, and labeling are followed. This involves storing all data in a secure, central location accessible to all team members; and storing all meta-data such as study protocols, consent forms, and data collection instruments such as interview guides, and documenting how these relate to the data (27, 31).

**Which Data and Materials Should I Share?** NIH’s policy defines the scientific data that must be shared as the “recorded factual material commonly accepted in the scientific community as of sufficient quality to validate and replicate research findings, regardless of whether the data are used to support scholarly publications” (1). As noted already, many in the QR community object to any notion of “replication” of findings (4). However, QR findings can be validated or verified: Someone can examine whether codes are reasonably applied to the data and whether theoretical claims find a basis in the data (12, 32). Moreover, a primary purpose of sharing data is to support high-quality secondary research, which would explore new questions with existing datasets (33, 34).

Given these aims, the primary form of QR data that must be shared are interview or focus group transcripts—the actual words of participants presented in a form that can be analyzed by others using standard techniques, including computer-assisted qualitative data analysis software (CAQDAS) (35). We believe that recordings are not the most appropriate form of data to share for two reasons: First, they identify research participants (voice prints and photographic images are HIPAA-safe harbor identifiers) (22); and second, recordings do not lend themselves to the most standard approaches to QR data analysis without duplicating transcription efforts.

Additionally, NIH policy requires the sharing of “accompanying metadata” which are defined as “Data that provide additional information intended to make scientific data interpretable and reusable” (1). In the context of QR, this would include a description of the study protocol sufficient to understand the sampling and data collection approach, the interview guide or similar data collection instrument, participant descriptors for each transcript (insofar as this is
Informed consent for data sharing can be approached as a regulatory issue or an ethical issue. From a regulatory perspective, sharing identifiable data ordinarily requires secondary users to apply to use data, to describe appropriate research purposes. The QDR similarly offers the requirement of secondary users to apply to use data, to describe appropriate research purposes. The QDR similarly offers the appropriate level of data security, protecting data from hackers, unauthorized users, or simple data loss. However, many repositories make all data open access, which may not be appropriate when data are sensitive or participant communities have requested or required restrictions on use. There may be times when open access is appropriate for QR data; for example, in oral history or interview projects where participants have chosen to be identified (39, 40), or when data are not sensitive and the risk of reidentification is low. In our pilot project, however, all data were deposited using the ICPSR's restricted access option. This requires secondary users to apply to use data, to describe how data will be protected, and to present an IRB-approved protocol (20). This allows repositories to offer a much higher level of protection and to ensure that data are only used for appropriate research purposes. The QDR similarly offers the option of access control, though the terms of access may differ widely depending on the nature of the data and the preferences of depositing researchers or participant communities (23). We expect that in the coming years many more repositories will develop capacity to curate, archive, and share QR data responsibly.

**Where Should I Share?** NIH encourages the use of established data repositories, rather than making data available directly from investigators or through institutional repositories (37). Working with an established repository offers several important advantages: Unique persistent identifiers for the dataset such as a DOI number, curation, clear processes for accessing data and guidelines for appropriate use, and state-of-the-art data security (23, 37).

In the United States, at least two repositories are currently equipped with the knowledge, experience, and policies to responsibly store and share QR data: the ICPSR at the University of Michigan and the QDR at Syracuse University. Both repositories have guidelines for preparing, depositing, and reusing qualitative data, and data curators with experience de-identifying qualitative data (or checking the quality deidentification) (23, 38).

We can reasonably expect that all data repositories will provide an appropriate level of data security, protecting data from hackers, unauthorized users, or simple data loss. However, many repositories make all data open access, which may not be appropriate when data are sensitive or participant communities have requested or required restrictions on use. There may be times when open access is appropriate for QR data; for example, in oral history or interview projects where participants have chosen to be identified (39, 40), or when data are not sensitive and the risk of reidentification is low. In our pilot project, however, all data were deposited using the ICPSR's restricted access option. This requires secondary users to apply to use data, to describe how data will be protected, and to present an IRB-approved protocol (20). This allows repositories to offer a much higher level of protection and to ensure that data are only used for appropriate research purposes. The QDR similarly offers the option of access control, though the terms of access may differ widely depending on the nature of the data and the preferences of depositing researchers or participant communities (23). We expect that in the coming years many more repositories will develop capacity to curate, archive, and share QR data responsibly.

**What Permissions Do I Need to Share Data and from Whom?** Informed consent for data sharing can be approached as a regulatory issue or an ethical issue. From a regulatory perspective, sharing identifiable data ordinarily requires informed consent, but sharing deidentified data does not, because federal regulations generally do not consider deidentified data to be human subjects data (45CFR46). From an ethics perspective, even when data will be deidentified, providing information on data sharing in the informed consent form is an important way of demonstrating respect for participants and being transparent about potential risks (5, 23). For prospective collection of data, consent forms should address data-sharing plans (38). The ICPSR and QDR repositories offer researchers template language for sharing qualitative data (41, 42). We believe that IRBs should offer some flexibility with template language to permit researchers to address important details: Will data be deidentified prior to sharing? Will data be open-access or restricted access? Will applications for secondary uses of data be reviewed by an IRB or other body? One-size-fits-all language is unlikely to provide potential participants with the information they require in the context of QR.

What if you want to share data already on the shelf, and the consent form did not explicitly disclose that data will be shared? Several factors determine whether data can nevertheless be shared. First, does the consent form explicitly forbid data sharing, or promise something incompatible with data sharing, such as, “all data will be destroyed”? In our experience, institutions will (rightly) prohibit data sharing when consent forms say such things. One option in such cases is to recontact participants to request permission for data sharing. This is, of course, an imperfect option given that it requires additional work and responses from former participants, contact information is often inaccurate as time lapses, and some participants may decline to share data. However, it may enable some important data to be shared. Prior efforts at recontacting participants have enabled data from a majority of participants to be shared (43), and some small studies indicate that most participants support sharing qualitative data when they are deidentified and shared only with other researchers (5, 19). This pattern has held even in studies of sensitive issues with participants who have survived trauma (44).

QR data sharing may also require institutional permission. In most cases, institutions that receive grants and employ researchers own research data (16). Researchers are generally viewed as stewards of the data for purposes of conducting research and are expected to initiate data sharing processes, but rarely do they own data or have complete rights to dispose of data as they see fit. The permissions needed for sharing QR data depend on what is being shared. Repositories often accept deidentified data that are not particularly sensitive without signatures or approvals from institutional officials. However, if data are not fully deidentified, are highly sensitive, or will be shared restricted access (e.g., due to sensitivity), institutional officials are usually involved in the deposit process. Researchers often assume that IRBs are the appropriate office to get involved. However, our pilot project found that data use agreements and data deposits are most commonly handled by an authorized official or institutional signatory from an office for sponsored projects or research contracts. The “common rule”—or federal regulations 45CFR46 on the Protection of Human Subjects—does not consider research with deidentified data to be human.
subjects research (45CFR46.102(e)); thus, sharing deidentified data falls outside of the purview of IRBs. In our experience, IRBs typically get involved only when data are not fully deidentified and contract offices have questions about whether consent forms permit data sharing.

Below, we address research with communities that may have additional requirements or procedures for obtaining permissions to conduct research or share data.

What Counts as Deidentifying Qualitative Data? Most data gathered from human participants in the United States is subject to the Common Rule (45CFR46), “Protection of Human Subjects.” The Common Rule states that research must make adequate provisions “to maintain the confidentiality of data” (46.111(7)) and the consent form should describe “the extent, if any, to which confidentiality of records identifying the subject will be maintained” (46.116.b(5)). However, the Common Rule does not specify how confidentiality must be protected nor does it provide any standards for the deidentification of data. Although the US HIPAA Privacy Rule does not govern most human research data, it does provide two commonly used standards for the deidentification of data. The “safe harbor” standard involves removing from a dataset 18 categories of information, including names, addresses (generally, more specific than states), dates other than years pertaining to the individual, phone and social security numbers, email addresses, photographic images, biometric identifiers including voiceprints, and any other unique identifier (45). The other standard is “expert determination.” Essentially, it requires that an individual familiar with “generally accepted statistical and scientific principles and methods” determine “that the risk is very small that the information could be used, alone, or in combination with other reasonably available information, by an anticipated recipient to identify an individual who is a subject of the information” (45).

These standards were developed with large, quantitative databases in mind (22); hence, reference to “statistical” principles and methods. Currently, no regulatory standards specifically reference the deidentification of qualitative data. We believe that the best approach in the current environment is to remove all HIPAA identifiers (thus satisfying a regulatory baseline requirement) and to use a nonstatistical variation of expert determination by examining texts for information that might lead to inferential identification of participants.

In addition to the HIPAA safe harbor variables, the QuaDS software that our project developed flags text that fall into several categories: locations more general than states, ages, dates and numbers, organizational names, race, ethnicity, and nationality, LGBTQI identity, and rare diseases (22). In flagging such a wide variety of variables, we are not proposing to expand the list of safe harbor variables, but rather to support “expert determination” efforts. That is, most text that is flagged can be safely left in research transcripts, but because these variables could lead to reidentification, they deserve review. Reviewers want to pay special attention to outlier traits such as a rare profession or a very high number of children, as well as rare combinations of traits (e.g., female drummer in a Serbian rock band). In our view, an expert determination need not treat every participant case alike; in one transcript, nationality might serve as an identifier, but not in another.

In some fields, where scientifically appropriate and feasible, the need to deidentify data may lead to more multi-site qualitative studies as a form of human subject protection; at a minimum, it should lead to greater efforts to mask study sites. At present, many QR studies are conducted at only one site (3). Knowing the site may reveal the geographic location of participants at a level much more specific than a state, and greatly facilitate reidentification of participants (34). Many traits that are relatively common in the United States may be rare in some localities.

We believe that the maxim, “as much as necessary, as little as possible” applies well to the deidentification of QR transcripts. “As much as necessary” highlights the need to remove information that creates a reasonable risk of reidentification; “as little as possible” highlights the need to leave in as many details as possible (while achieving deidentification) to preserve a sense of context and human narrative.

While we often refer to “removing” information, in fact, in most cases, information is best replaced or masked by changing specific information to more general information, for example, “University of Kansas” to “Midwestern University” or “4th grade reading teacher” to “elementary school teacher.” When masking, it is essential to have rigorous protocols to track changes and ensure consistent substitutions for the same information (e.g., if Robert is changed to James, then this change must be made consistently throughout a transcript). The QuaDS software automates this process, allows for customized changes, and generates change logs. This can also be done manually.

The expert determination standard requires removing, masking, or rearranging information until “the risk is very small” of reidentifying a participant with retained information (46); this suggests that deidentification is always a matter of degree. When determining how much information to remove, researchers will want to consider the consequences of reidentification: Are the data particularly sensitive? Would reidentification cause social, emotional, legal, or financial harm? In such cases, we believe it best to err on the side of restricting access in some fashion. There are many different levels of restricted access, but most restrict access to researchers who have appointments at universities and IRB approval to conduct secondary research.

QR data are often collected as part of mixed-method studies that involve gathering quantitative data including numerous demographic datapoints (i.e., potential identifiers) (35). Naturally, the risk of reidentifying participants using QR transcripts increases when potential identifiers in QR transcripts can be triangulated with a more detailed set of participant descriptors. Mixed-methods datasets will thus require special consideration, increasing the likelihood of restricting access or removing variables one might have left were qualitative and quantitative datasets not shared in tandem.

Finally, it is worth acknowledging that in some cases an individual participant might discuss important things that are too unique to be deidentified adequately. In such cases, NIH guidance recommends access controls (2). Depending on the risk level, researchers might also request permission to withhold a specific transcript from the deposited dataset.
How Can I Support High-Quality Secondary Research with My Data? A 2021 special issue of The Qualitative Report included seven articles that each analyzed the same dataset; the purpose was to illustrate the vast array of approaches to analyzing the same data (47). Interest in secondary analysis of QR data has grown significantly over the past two decades (48–51). However, one of the most common objections to QR data sharing is that secondary users will lack the context and relationships necessary to interpret data (52, 53). We offer several replies.

First, in many cases, qualitative researchers do not actually have substantial relationships with their participants. There are clear exceptions; for example, in some ethnographic studies, researchers live in a community that is being studied (54); and in some community-based participatory research studies, the line between researchers and participants is often blurred, with members of communities playing important research roles (55). However, in our recent review of 100 QR articles published in health-related journals, the vast majority of articles involved a one-time contact with participants that lasted 1 h or less, and fewer than half of article authors were involved in data collection (3).

Second, there may be some advantages to reading the words of participants with fresh eyes. While there is a risk that a secondary researcher may misunderstand something, there is also the possibility that the researcher may perceive something new that is valid or ask new questions that can be legitimately explored using the dataset. As Kuula wrote over a decade ago:

... the idea that the original researcher is the only one capable of analyzing the data correctly means that the original methodological is the orthodox way to understand research data. What this implies is that the original researcher has an exclusive right to define the characteristics and nature of the empirical world under investigation. That is an odd presupposition for a research paradigm that often accuses quantitative re-search of naive realist epistemology (43).

Third, and most importantly, there are many things researchers can do to assist secondary researchers in understanding their study and their data (23, 38). Transcripts should be accompanied by any participant descriptors that were included in data analysis. Data must be deposited with data collection instruments such as interview guides and citations or links to relevant publications on the data. If some details about the participants or setting are crucial to understanding the data, they should be provided in background documents. If a researcher still feels their perspective and experience are essential to interpreting data, they can post a note with their data stating this and offering collaboration or consultation (36).

Finally, asking the repository to restrict access may support higher-quality research in at least two ways. First, secondary users must agree not to attempt reidentification of participants and to protect the confidentiality of data. This means that more participant descriptors might be responsibly retained. Second, restricted access data often can be accessed only by individuals with academic appointments and IRB or research ethics committee approval. This may help ensure a baseline of competence on the part of the secondary researcher, making it more likely that limitations to their approach will be disclosed in publications.

How Can I Respect the Interests of Community Partners? It is often challenging to determine when research occurs with a “community” versus a population; few “target populations” are structured formally (56). Further, many communities do not speak with one voice—their members may have competing priorities (57). Nevertheless, when communities of participants exist, they may want some control over how data are used—what questions are explored, whether or how comparisons are made to other groups, and who interprets and publishes the data (58–60). Researchers have earned the mistrust of some communities by failing to offer standard protections, including appropriate informed consent, by publishing findings in ways that stigmatize them, and by prioritizing their interests over those of the participant communities (59–62). Many qualitative researchers acknowledge the mistrust that some groups have regarding widespread data sharing (15, 21).

Open-access data sharing removes any control that communities might have over how their data are used. While US regulations typically view data as belonging to the institution that receives federal funding for research, some communities do not view it this way, and in research on Tribal lands, Tribal law applies.

NIH has provided supplemental guidance specifically on “Responsible Management and Sharing of American Indian/Alaska Native Participant Data” (63). This document emphasizes some principles that are appropriate for any community-based research project, such as proactive engagement on data stewardship, uses, and sharing; establishing mutual understanding; and protecting confidentiality. However, some elements of Tribal research are unique, including the possibility that the data repository that is used might exist on Tribal lands and be subject to Tribal law.

The “CARE Principles for Indigenous Data Governance” provides an excellent framework for data sharing in the context of community-engaged research. The CARE principles emerged from a 2018 international workshop in Botswana led by Indigenous peoples with academics and informatics practitioners, which drew heavily upon prior work by national and international teams (64). The acronym CARE represents the principles of Collective benefit, Authority to control, Responsibility, and Ethics. While the FAIR principles are “data-centric,” the CARE principles are more strongly focused on people and purposes (64). While these principles are generally easily reconciled with a strong commitment to data sharing, they may conflict specifically with default “open-access” practices (64). Sharing data with restricted access, and negotiating the terms of access, is a promising strategy for partnering with communities to share data.

While it is understandable that some communities may have concerns with data sharing, there are also reasons for communities to welcome it. Data sharing can be empowering. In principle, it provides community partners with access to complete datasets and opens up the possibility of analyzing data with a focus on questions of particular interest to the community. In our initial article, “Is it time to share qualitative data?” we noted how qualitative researchers necessarily play a major role in filtering
findings, determining what is considered a theme and which quotes are key. We noted that in one of our own focus group studies, whose findings were published in a major medical journal, we were able to share “one quote for every 45 min of conversation with a highly engaged group”... 80% of participants “never had any of their words shared in print” (8), p. 1. Sharing QR data enables participants and their communities to see complete findings, to identify major themes from discussions for themselves, and to challenge or supplement published perspectives, which are necessarily incomplete.

When Should I Share Data, Deidentify Data, and Destroy Identifiable Data? The new NIH data sharing rules and White House guidance require data sharing at the time of first publication or sooner, and embargos are not allowed (1, 7). No longer can data sharing be envisioned as a task to tackle once a complex project is complete. Consider a study that involves a survey with some open-ended items followed by in-depth interviews aimed at understanding more deeply survey findings. The study concludes with a consensus process—a Delphi panel—to develop recommendations informed by the study’s primary findings. Data or project materials may need to be deposited at four or more timepoints: as each component of the study is published (e.g., the survey, the interviews, and the Delphi) with a final deposit of outstanding materials for the entire project, including publications.

Compliance with NIH requirements might be accomplished piecemeal, with some data shared through a journal as online supplemental material and some deposited with a repository. However, the final deposit will then either lack the full set of data or will create redundant records. Such an approach risks violating the “findability” and “reusability” FAIR principles.

We strongly recommend selecting one repository during the project planning phase and communicating with the repository early in the life of the project to ensure that all deposited data are clearly designated as part of the same project. This will enhance data findability and completeness and avoid duplication of datasets and DOIs (65).

We also recommend “cleaning while you cook”: deidentify data either prior to coding or immediately after coding. There are several advantages of working with deidentified data: You are likely to use meaningful replacements for original text that needs to be masked, because your key or illustrative quotes will use this text; it enhances confidentiality protections throughout the project; and it reduces the temptation to rush the job of deidentification prior to publication. As this recommendation illustrates, the new policy will require most qualitative researchers to fundamentally change the way they work with data.

In general, we discourage destruction of original data files such as audio or video recordings, even though these may be retained only on institutional servers for purposes of confidentiality. First, retaining data may be necessary to comply with the requirements of funding agencies or professional ethics codes (66). Second, retaining original data may be the only way to protect oneself if allegations of data fabrication or falsification are made, and these are frequently made long after minimum data retention requirements have been fulfilled (67).

Are We Ready for Qualitative Data Sharing?
The questions we briefly engaged above address important issues pertaining to data usefulness, participant trust and protection, and legal and logistical matters. In our interviews with qualitative researchers, we found a lack of knowledge of very basic elements of QR data sharing. This is not surprising given that nearly all qualitative researchers lack experience sharing data. More concerning is the fact that guidance has been lacking even for those who seek it. One would expect to find guidance from at least three sources: data repositories, QR journals, and textbooks on QR methods. As we prepare to disseminate a QDS Toolkit, we investigated existing guidance and opportunities to share information and resources on QDS. For this purpose of preparing for dissemination, we conducted reviews of guidance from data repositories, journals, and textbooks.

Data Repositories Serving the Social Sciences. In 2015, our team conducted content analysis of English-language guidelines from global repositories that accept social science data or multidisciplinary data (6). As our current study focuses on preparedness for QDS in the United States, our team reviewed the guidelines from the US repositories from our 2015 search and repeated our search to see if any additional US repositories now have publicly available guidance on QDS.

In 2015, we included five US repositories in our analysis: The Murray Research Archive at Harvard; ICPSR at the University of Michigan; QDR at Syracuse; Databrary; and the Dryad Digital Repository. In the current search, we found an additional 2 US repositories that accept social science or multidisciplinary QR data (the CISER Data & Reproduction Archive at Cornell and the Data Archive & Collections at University of California, Los Angeles). Of the seven repositories, only two have publicly available guidelines specific to QR data (ICPSR and QDR) (41, 42). See SI Appendix, Table S1 for full results.

QR Journals. In April of 2021, we systematically searched the National Library of Medicine catalog to identify academic journals dedicated to publishing primarily qualitative or mixed-methods data on health science topics. To be included, journals had to be available in the United States, publish peer-reviewed articles in English, and be focused on presenting qualitative or mixed-methods data on health science topics; and they could not be identified as a predatory publisher on Beall’s List. The final list resulted in 14 journals included in the dataset.

For each included journal, we identified the author submission guidelines on the journal’s website, saved them as a PDF, and uploaded them to Dedoose qualitative data analysis software to be coded for references to data sharing. One of three codes was applied to each set of guidelines: 1) requires data sharing; 2) encourages but does not require data sharing; or 3) silent on data sharing, indicating no direct references to data-sharing policies. Of the 14 journals, slightly more than half were silent on data sharing (n = 8), while the remaining journals encouraged it (n = 6); one of these latter journals requires that data be provided if requested by a competent professional who seeks to verify substantive claims under special conditions. Journal titles and full results are reported in SI Appendix, Table S2.
QR Textbooks. In May of 2021, the research team identified and reviewed the analytic table of contents for textbooks listed on Amazon.com designed to teach QR methods. In order to meet inclusion criteria, the textbook had to have been listed in the Amazon marketplace as a book in “new condition,” written in English, and published between 2016 and 2020. The search term “qualitative research methods textbook” returned 769 results. Each textbook’s table of contents preview link was reviewed to ensure that it met inclusion criteria; at least 50% of the table of contents needed to be focused on QR methodology with human subjects—for example, references to “qualitative chemistry” were excluded. For textbooks with multiple editions available, only the most recent edition was included in the present dataset. These two factors resulted in the exclusion of 655 search results. Textbooks also needed to provide a detailed analytic table of contents to be included; 77 textbooks were excluded due to this criterion. The final dataset included 41 qualitative methodology textbooks. Only three of these textbooks on QR methods dedicated a chapter or chapter subheading to data sharing, and all focused on coding data using a specific software package. (See SI Appendix, Table S3 for a list of textbooks and detailed review results.)

As we write this (June 2023), the first cycle of NIH grants submitted under the new policy has just been reviewed; some will soon begin gathering data, and as they publish, data will need to be shared. When we first launched our project in 2017, few resources existed, and few people had experience sharing QR data. Our reviews of data repositories, qualitative health research journals, and QR textbooks indicate that guidance on QR data sharing is still not widespread, and our surveys indicated a general lack of knowledge about QR data sharing on the part of researchers, IRBs, and even repositories. However, much has improved since our project first began. The QDR at Syracuse and the ICPSR at Michigan have gained experience curating human participant interview data and have produced or updated guidance; NIH has produced a series of guidelines to help researchers navigate a series of challenges associated with QR data sharing; studies of participants indicate broad support for sharing QR data with appropriate protections; and the QuaDS software makes data deidentification significantly easier and faster than manual deidentification.

Beyond Logistics and Preparedness

As with most goals in research, our goals with QR data sharing cannot be met by adopting a mindset of satisfying a requirement. Mandating data sharing has advantages, but without appropriate oversight and buy-in from researchers, it could be problematic. Two risks are particularly concerning: the potential that participants will be reidentified and harmed by the confidentiality violation and the risk that deposited data will not meet the basic requirements of FAIR principles for data sharing, thereby thwarting the fundamental purposes of data sharing.

Risk of Participant Reidentification. To our knowledge, there are no mechanisms currently in place to prevent (rather than prohibit or discourage) a researcher from depositing data open access with no data curation or proper data deidentification; and most open access datasets can be downloaded by anyone, with no enforcement of secondary user purposes or data protections. We hope that researchers and secondary users will always behave responsibly, but we also expect that data sharing will require more oversight than currently exists to ensure proper data curation and protection. We admit to some level of worry about open-access data repositories that will accept data with no curation, no institutional approval, and no review of data sensitivity or deidentification.

Oversight of data sharing might occur at several levels. Those who receive data—repositories, and most especially open-access data repositories—may have an obligation to ensure that data are appropriately deidentified. This is important because open-access deposits frequently can be made without the approval of institutional officials, and depositing researchers may be in a hurry to comply with a mandate so they can publish findings they have worked long and hard to prepare in publishable form. People do not routinely read the fine print of agreements when using online services; so it is of little consolation that these repositories might require an individual who is in a hurry to deposit data—or an individual keen to download data—to click a box agreeing to terms of use (68).

Institutions may have an obligation to create data-sharing policies—just as they have all created conflict of interest and other policies—which clarify who has the authority to deposit data, where, when, and under which conditions. This is important because institutions—e.g., universities that receive research funding—typically are the owners of data (16, 66). It is important to note that deidentified data are not typically in the purview of institutional review boards (IRBs) but identifiable data are. When data are not appropriately deidentified, a whole range of new obligations arise that need to be addressed by policies.

Risk that Deposited Data Are Essentially Useless. Those who mandate data sharing (e.g., journals and funding agencies) may have a responsibility to ensure that the data deposit are in fact of sufficient quality to support verification of findings and secondary data analysis. A deposit without full transcripts will not permit either verification of findings and secondary data analysis. A deposit made without the approval of institutional officials, and those who receive data—repositories, and most especially open-access data repositories—may have an obligation to ensure that data are appropriate.

Institutions may have an obligation to create data-sharing policies—just as they have all created conflict of interest and other policies—which clarify who has the authority to deposit data, where, when, and under which conditions. This is important because institutions—e.g., universities that receive research funding—typically are the owners of data (16, 66). It is important to note that deidentified data are not typically in the purview of institutional review boards (IRBs) but identifiable data are. When data are not appropriately deidentified, a whole range of new obligations arise that need to be addressed by policies.

Institutions may have an obligation to create data-sharing policies—just as they have all created conflict of interest and other policies—which clarify who has the authority to deposit data, where, when, and under which conditions. This is important because institutions—e.g., universities that receive research funding—typically are the owners of data (16, 66). It is important to note that deidentified data are not typically in the purview of institutional review boards (IRBs) but identifiable data are. When data are not appropriately deidentified, a whole range of new obligations arise that need to be addressed by policies.

Institutions may have an obligation to create data-sharing policies—just as they have all created conflict of interest and other policies—which clarify who has the authority to deposit data, where, when, and under which conditions. This is important because institutions—e.g., universities that receive research funding—typically are the owners of data (16, 66). It is important to note that deidentified data are not typically in the purview of institutional review boards (IRBs) but identifiable data are. When data are not appropriately deidentified, a whole range of new obligations arise that need to be addressed by policies.

Institutions may have an obligation to create data-sharing policies—just as they have all created conflict of interest and other policies—which clarify who has the authority to deposit data, where, when, and under which conditions. This is important because institutions—e.g., universities that receive research funding—typically are the owners of data (16, 66). It is important to note that deidentified data are not typically in the purview of institutional review boards (IRBs) but identifiable data are. When data are not appropriately deidentified, a whole range of new obligations arise that need to be addressed by policies.

Institutions may have an obligation to create data-sharing policies—just as they have all created conflict of interest and other policies—which clarify who has the authority to deposit data, where, when, and under which conditions. This is important because institutions—e.g., universities that receive research funding—typically are the owners of data (16, 66). It is important to note that deidentified data are not typically in the purview of institutional review boards (IRBs) but identifiable data are. When data are not appropriately deidentified, a whole range of new obligations arise that need to be addressed by policies.
Data sharing has immense potential to advance transparency and scientific rigor in QR, advance knowledge by increasing the possibilities realized, and data repositories.

ACKNOWLEDGMENTS. We thank the three reviewers and article editor for many helpful suggestions, including several suggestions for new references.

This paper was funded by the US National Human Genome Research Institute (R01HG009351).