ICPSR

Inter-university Consortium for Political and Social Research

Institute for Social Research
P.O. Box 1248
Ann Arbor, MI 48106-1248

ICPSR-help@umich.edu
(734) 647-2200

www.icpsr.umich.edu

The Inter-university Consortium for Political and Social Research (ICPSR) is pleased to provide this response to the National Institute of Health's Request for Public Comments on Supplemental Information to the NIH Policy for Data Management and Sharing: Protecting Privacy When Sharing Human Research Participant Data. ICPSR collects, curates, and disseminates data covering a broad spectrum of disciplines, including political science, sociology, demography, economics, education, psychology, criminology, gerontology, public health, public policy, and more. ICPSR data form the foundation for tens of thousands of research articles, reports, and books that advance science.

The collection includes over 17,000 data studies, including almost 2000 restricted data studies. In addition to distributing data supported by its members, ICPSR partners with more than 20 government agencies and foundations, including the Bureau of Justice Statistics, the National Institute on Aging, the Eunice Kennedy Shriver National Institute of Child Health and Development, the Bill & Melinda Gates Foundation, and the Robert Wood Johnson Foundation, to disseminate their data, thus fulfilling mandates to make data publicly available. The special collections funded by agencies and foundations provide researchers and practitioners in these fields with key data resources as well as customized tools to support their work.

Recognizing that the increasing complexity of data requires new levels of support and guidance in their proper use, ICPSR has developed expertise in facilitating access to and secondary analysis of complex data – e.g., longitudinal data, linked data, geospatial data, video data, and data containing biomarkers. The organization distributes enhanced data and documentation to researchers in forms that facilitate reuse of complex data and offers training and consultation. ICPSR currently provides advice to researchers on protecting the confidentiality of data while making it accessible for secondary analysis. ICPSR provides multiple, tiered modes of access to its restricted data.

ICPSR applauds NIH's efforts to provide thoughtful and concrete guidance to the research community to ensure responsible implementation of its new policy for data management and sharing. We strongly support the principle underlying this guidance, namely that it is possible and ethical to protect and respect the privacy of data subjects *and* to support increased transparency, reproducibility, efficiency, and knowledge creation in the scientific enterprise through responsible data sharing. We also urge NIH to acknowledge that responsible management of data, including de-identification, takes resources and therefore to provide support to both investigators and data repositories so that they undertake the necessary work.

We think it is important that de-identification can undermine data quality and fitness for use. It is important for investigators and IRBs to understand that it is possible, ethical, and scientifically appropriate to share unsafe data in safe ways by making the data available in safe places, restricting use to safe researchers, and ensuring that the resulting research produces safe output. Encouraging investigators and IRBs to make use of the Five Safes framework would help them to find appropriate methods to maintain confidentiality of data subjects while advancing scientific research.

ICPSR encourages investigators to include restrictions on the use and sensitivity of data in machine actionable, standardized metadata, so that access and protection can be automated. Data that is not unsafe should be easily accessible. There may still be licensing, requirements to cite, not re-disseminate, registration, and payment of reasonable fees. Maintaining access to data is not costless, and cost recovery is consistent with data sharing and "open" science. Investigators should be encouraged to work with certified repositories that know how to protect data and provide secure access. This would significantly increase safe, responsible access to secondary data relative to a strategy in which each institution receiving NIH funding is expected to have expertise in disclosure review and providing access to restricted data.

We offer specific suggestions below in hopes that they will help NIH to strengthen and clarify its guidance.

**ICPSR comments on DRAFT operational principles:**

> 1. NIH and the institutions it funds are obligated and required to protect the privacy and confidentiality of every participant as described in informed consent and in line with all applicable laws, policies, and regulations.

In order to support researchers in following this principle, ICPSR recommends that NIH provide additional guidance. Researchers and research institutions may well be unaware of or confused by relevant laws and regulations. Concern that they might inadvertently violate relevant laws or regulations may discourage data sharing (or provide a pretext for refusing to share data). We recommend that NIH provide guidance that specifically addresses key legislation or regulations that might raise concerns about sharing data. The U.S. Department of Health and Human Services, Department of Administration for Children and Families (HHS/ACF) has contracted with Westat, ICPSR, and The Future of Privacy Forum (*ACF Privacy and Confidentiality Analysis Support*, 75P00120F37018) to create a database of

legislation and regulations that intersect with data privacy and data sharing that are relevant to ACF programs and child welfare-related organizations. The goal is to create an accessible resource for ACF staff with minimal legal background to understand what federal, state, and tribal laws say about data sharing and privacy for both human services programmatic and research purposes. Currently, the database covers over 60 relevant federal laws. We recommend that NIH leverage this valuable resource and expand it to encompass legislation and regulations that affect NIH researchers. See the Confidentiality Toolkit for an example of a product that has been created from this database.

> 2. Researchers and institutions should proactively assess appropriate protections for sharing scientific data from participants, including determining whether sharing should be restricted through controlled access,[4] regardless of whether the data meet technical and/or legal definitions of "de-identified" and can legally be shared without additional protections (e.g., the research does not meet the definition of "human subjects research" under the Common Rule).

ICPSR strongly supports NIH's recommendation that researchers and institutions address these issues proactively. We encourage researchers to *use* their data management and sharing plans to manage their data throughout the research data lifecycle. As indicated in NIH's guidance regarding consent statements, it is particularly important that researchers and institutions begin planning for data sharing prior to beginning data collection. Early planning ensures that researchers design data collection to minimize the collection of data that could compromise confidentiality, to establish systems to protect data with appropriate IT security, staff training, and the separation and/or encryption of direct identifiers. We also recommend that research teams identify and reach out to an appropriate long-term home for the data at the beginning of the project; repositories are important sources of information and guidance, and can assist researchers in assuring that the data meet the requirements of the expected host repository. Early planning also makes the creation of documentation and appropriate metadata much easier, facilitating reuse. Inclusion of any restrictions on use can also be built directly into machine-readable and ideally machine-actionable metadata.

> 3. Investigators and institutions should develop robust consent processes that prioritize clarity regarding future sharing and use of scientific data, including limitations on future use, and general aspects regarding how data will be managed (see Informed Consent for Secondary Research with Data and Biospecimens: Points to Consider and Sample Language for Future Use and/or Sharing).[5] Importantly, when a study offers the possibility of a direct benefit for research participants, the DMS Policy does not require sharing of data in order to participate.

We strongly support NIH's recommendation that investigators and institutions develop robust consent processes that prioritize data sharing and re-use of scientific data. We also strongly support the underlying commitment to ethical treatment of

research participants and their data that motivates the option for participants to refrain from sharing their data while participating in potentially beneficial research. Consistent with this ethical use of data, we support research that is intentionally inclusive and ensures that marginalized groups are well represented in scientific data. Recruitment of and communication with participants should emphasize the benefits of and the importance of inclusion and representation in secondary data. Making people invisible does not protect them.

> 5. Collection of data from non-traditional research settings, such as mobile health devices, social media, consumer reports, and public health surveillance also warrant strict privacy considerations.

We recommend adding language to this statement to clarify that NIH is not discouraging people from using and sharing these non-traditional data sources in appropriate, privacy-protecting ways as they have significant scientific potential. For example:

> *5. Collection of data from non-traditional research settings, such as mobile health devices, social media, consumer reports, and public health surveillance has enormous scientific potential. Sharing these data may warrant specific, appropriate privacy considerations.*

## ICPSR comments on DRAFT best practices:

> 1. Ensure Appropriate De-identification. NIH recommends scientific data to be de-identified to the greatest extent possible in a manner that maintains sufficient scientific utility.

Defining "appropriate de-identification" is perhaps the most challenging issue raised in this guidance. The appropriate level of de-identification depends on the mode of data access and the credentials of the user. Where access is controlled, and the ability to combine sensitive data with other data that would increase the risk of re-identification is limited, when the user has significant experience with and understanding of confidentiality protection, when there are credible incentives to maintain confidentiality -- "appropriate" de-identification may be very different than where the data is freely accessible and can be combined with publicly available data that has direct identifier. We recommend a change in wording to avoid suggesting that protecting privacy by removing potentially identifying data should trump scientific utility when there are other ways to protect confidentiality.

> *1. Ensure Appropriate De-identification. NIH recommends scientific data be de-identified and managed as necessary to protect privacy and maintain scientific utility.*

Investigators and institutions should learn and make use of best practices for safely sharing data that is potentially re-identifiable. For many important research questions, de-identification significantly undermines the scientific utility of data.

Restricted data use agreements, controlled computing spaces (e.g., virtual and physical data enclaves), and the separation and encryption (but not destruction) of direct identifiers can all facilitate scientific analyses on data that is re-identifiable while safely protecting the confidentiality of data subjects. Researchers should generally not try to disseminate restricted data themselves, but should work with a trusted repository that has experience in this area.

It may help users if NIH explicitly defines direct and indirect identifiers and connects these definitions to the implications for sharing data and determining appropriate access restrictions. ICPSR defines direct identifiers as variables that point explicitly to particular individuals or units, including names, personal addresses, full dates, unique identifying numbers (i.e., SSN), telephone/fax numbers, email addresses, URLs that only allow restricted access, IP numbers, biometric identifiers (i.e., fingerprints), full face or similar photo images, and X/Y GPS coordinates. We distinguish these from indirect identifiers that may identify an individual or unit in combination with other variables. Examples include: race, ethnicity, demographics, income, medical history, etc. In creating public use data products (i.e., those without restrictions on use), we always remove direct identifiers. We then consider whether there is sensitive subject matter (and hence greater risk of harm) or indirect identifiers that increase the risk of inferential disclosure. Either of these criteria may result in a data set being classed as restricted-use. Any data with direct identifiers is restricted use, and direct identifiers are separated from other data and stored with an encrypted crosswalk. In general, qualitative data is presumptively restricted, as de-identifying it is generally complex and costly and would lead to significant deterioration in scientific value. This is also the case for much video data.

> 2. Establish Scientific Data Sharing and Use Agreements. NIH recommends the use of scientific data sharing and/or use agreements, preferably standardized, when sharing data from participants with and from repositories. These agreements should be considered even if scientific data are de-identified[11] and should be negotiated among researchers, institutions, and repositories.

We strongly support the use of data sharing and use agreements. Formal data use agreements, with institutional signoff, support the creation of a culture of confidentiality and accountability. We recommend that investigators be encouraged to produce standardized, machine actionable metadata that captures the terms of data use agreements including any restrictions on use of the data, such as those promised in participant consent statements. This will ensure that the terms promised to participants are carried through by subsequent, secondary users of the

data, and reduces delays in access that can be associated with the use of formal agreements.

> Oversight. Agreements should clearly include certification from an institutional official that, at a minimum, scientific data have been appropriately de-identified (and to which standard), that an institutional oversight body has reviewed and considered the risks of data sharing, and that sharing is consistent with informed consent (as applicable).

We are concerned about potential interpretations of this paragraph on Oversight. The phrasing here suggests that these institutional officials (presumably university lawyers or an IRB) would have to certify that data were de-identified before they could be shared. Many institutions do not have institutional officials who can provide such certification. The expertise to do this is often housed in the research investigators themselves, or in a trusted, certified repository to which the data are headed. It would make more sense, and result in less diminution of scientific value, to have certification conducted elsewhere, and the certification should be that the data are being shared safely, not that they are de-identified. NIH should invest in institutions that can provide this expertise and guidance to individual investigators.

We do not want to put a responsibility on IRBs that will simply encourage them to prohibit data sharing. IRBs are not as resistant to data sharing as they have been in the past (see, for example, Dessislava Kirilova, Diana Kapiszewski, & Colin Elman "Optimizing Openness in Human Subjects Research: Balancing Transparency and Protection" Global Virtual Conference (IASSIST 2021). Zenodo. https://doi.org/10.5281/zenodo.5181109. NIH should support continued education of IRBs to ensure that IRBs understand that researchers have an ethical responsibility to participants in scientific research to generate the most value possible from the data generated by their participation in research. This ethical responsibility is fulfilled by conducting transparent and reproducible research and facilitating safe secondary use of scientific data.

We recommend revision of the statement that data sharing agreements should include "certification from an institution official that, at a minimum, scientific data have been appropriately de-identified." For example, the following wording addresses the goals of confidentiality protection more effectively by allowing the investigators and the IRB to identify an appropriate agent to manage any de-identification or restricted sharing of data.

> *Oversight. Agreements should clearly include certification from an institutional official that an institutional oversight body has reviewed and considered the risks of data sharing and approved a plan for sharing that is consistent with the investigators' ethical commitment to protect the confidentiality of data subjects and with informed consent (as applicable).*

**ICPSR comments on DRAFT Points to Consider for Designating Scientific Data for Controlled Access:**

In order to encourage the broadest possible access to scientific data, the general point should be made that, if there is no risk of re-identification, the data are not sensitive, and no promise to restrict access has been made, the data do not need to have controlled access. We would rephrase the following sentence:

> In cases where participants explicitly consent to share scientific data without restrictions, it may be appropriate to share data without access controls.

The slight change in the wording in the sentence below assures investigators and IRBs that making safe, de-identified data available without restrictions is perfectly acceptable and ethical.

> *In cases where participants explicitly consent to share scientific data without restrictions, it is generally appropriate to share data without access controls.*

Even where there are no concerns about confidentiality that require controlled access, it is still perfectly legitimate for secondary users to be required to accept terms of a license to use the data. For example, such terms of use may specify that they will use the data responsibly, will not try to re-identify any data, and will cite the original data producer. Data can be made available using a number of different licenses, such as Creative Commons licenses, Open Data Commons licenses, or a Public Domain Mark, and these should be distinguished from controlled access as described above. These licenses may address permission to re-share data (or not) and requirements to cite any use of the data. The use of such licenses establish an ecosystem that supports data sharing, acknowledgement of data use, and maintenance of data provenance and quality (which can be diminished with ad hoc or viral data sharing). For example, in addition to agreeing to comply with the applied open licenses, users of ICPSR's self-publishing repository agree: 1) Not to use the datasets for investigation of specific research subjects, except when identification is authorized in writing by ICPSR; and 2) To make no use of the identity of any research subject discovered inadvertently, and to advise ICPSR of any such discovery. ICPSR provides more guidance about licensing alternatives for data sharing at https://www.openicpsr.org/openicpsr/faqs.

With respect to point [2]

> Investigators should consider sharing participants' scientific data through controlled access repositories if data ….
>
> 2. Could be considered sensitive, such as including information regarding potentially stigmatizing traits, illegal behaviors, or other information that could be perceived as causing group harm or used for discriminatory

> purposes. Sensitive data may also include data from individuals, groups, or populations with unique attributes that increase the risk of re-identification.

We completely agree that it may be appropriate to restrict data due to potential harm to individuals or groups. However, it is important that this kind of "protection" not impede scientific research on marginalized populations. Making marginalized groups invisible is not responsible or ethical. Communities should have access to data about themselves, and may reasonably expect to have a voice regarding others' access to data about them.

Finally, we strongly agree with the point made in [3]:

> Investigators should consider sharing participants' scientific data through controlled access repositories if data ….
>
> 3. Cannot be de-identified to established standards or cannot sufficiently reduce the possibility of re-identification. Access controls, among other measures, may be appropriate to further mitigate the risk of re-identification.[16]

When data cannot be sufficiently de-identified without the data's fitness for use, the data should be available for research using controlled access methods. Those methods should use both technical and administrative methods to ensure confidentiality protection. Repositories with experience in sharing restricted data are recommended to facilitate both confidentiality protection and researcher access to such data.