Journal of Privacy and Confidentiality
Vol. 13 (2) 2023

NAHDAP-ICPSR

Submitted Nov 2022
Published Nov 2023

NATIONAL ADDICTION AND HIV DATA ARCHIVE PROGRAM: DEVELOPING AN APPROACH FOR REUSE OF SENSITIVE AND CONFIDENTIAL DATA

KATHY ETZ[†], HEATHER L. KIMMEL[†], AND AMY PIENTA[‡,*]

[†] National Institute on Drug Abuse, National Institutes of Health

[‡] ICPSR, University of Michigan

Abstract. Sharing data produced through health research projects has been increasingly recognized as a way to advance science more rapidly by facilitating discovery and increasing rigor and reproducibility. Much of the data collected from human subjects includes sufficient sociodemographic detail and/or covers sensitive topics, and thus requires restricted data management and sharing practices. Over the last two decades, scientific organizations, presidential memoranda, and other sources have all called for increasing opportunities to share data. Recognizing the value of shared data, the National Institutes of Health issued a new Data Management and Sharing Policy, effective January 25, 2023. Prior to this updated policy, in 2009, the National Institute on Drug Abuse recognized the value of sharing data and established an archive, the National Addiction and HIV Data Archive Program. This program focused on sharing data, often highly sensitive, generated from social and behavioral addiction research, including quantitative and qualitative assessments as well as biomarker and imaging data. NAHDAP has developed practices and curation standards to ensure datasets are improved and usable, and provides technical assistance for both data depositors and users. We share three key lessons learned working to disseminate sensitive data over the last 13 years, including (1) protecting the confidentiality of human subjects; (2) ensuring careful consideration of costs for archiving data requiring protection ; and (3) providing support to facilitate the discovery and use of the data.

1. Introduction

Over the last two decades, there has been increasing emphasis on making health research data more broadly available by setting an expectation for researchers and data producers to share data. At the same time, protecting the privacy of human subjects' data remains one of the

Key words and phrases: Confidential Data; Data Reuse, Sensitive Data;

The views and opinions in this report are those of the authors and do not necessarily represent the official position of the U.S. Department of Health and Human Services or any of its affiliated institutions or agencies. Dr. Etz and Dr. Kimmel were substantially involved in the scientific management of and providing scientific expertise for contract #75N95019C00017.

[*]Corresponding author, apienta@umich.edu.

www.journalprivacyconfidentiality.org
DOI:10.29012/jpc.853

most important concerns of clinical and health researchers regarding data sharing (Federer, et al., 2015) and an important consideration for designing data repository infrastructure.

1.1. **Why Share Confidential Data?** Data sharing, with appropriate protections in place, is useful to support research transparency and analytic reproducibility, to generate new findings, to inform the development of new measures and methods, and for training and education purposes. Data utility can also be increased by combining with (e.g., harmonizing or integrating) and linking to other datasets (e.g., administrative data and data from other disciplines) (Lohr and Raghunathan, 2017), expanding the research questions that can be addressed. Sharing data ensures that investments in data collection have the potential to be fully leveraged by facilitating more analyses of a broader range of questions than could likely be completed by the original research team. Shared data can also be a relatively low-cost way to provide pilot data for additional research and otherwise serve to advance the field. This set of goals remains important for data covering sensitive content areas, albeit how data sharing solutions are built must consider management of human subject protections before reuse is possible.

1.2. **National Addiction and HIV Data Archive Program (NAHDAP).** In 2009, the National Institute of Drug Abuse (NIDA)established a data archive (a term used synonymously with data repository) within the Inter-university Consortium for Political and Social Research (ICPSR) in response to increasing recognition of the utility of shared data and to be on the cutting edge of data sharing. The National Addiction and HIV Data Archive Program (NAHDAP) data archive focuses on archiving nationally representative and regional data from social and behavioral studies of substance misuse, including longitudinal data and repeated cross-sectional data that can be used to study behavioral changes over time among individuals and the U.S. population. Importantly, NAHDAP was conceptualized as a full-service archive to ensure the greatest utility and benefit to addiction and HIV science, hence the inclusion of "Archive Program" in the name. This signaled that the archive does more than simply preserve data. For NIDA, broad access to social and behavioral science substance misuse data was an important way to leverage and extend the reach of federal research investments alongside the goal of protecting sensitive and/or detailed attributes of studies and human subjects' information contained within the data.

As addiction is a highly sensitive topic, the archive had to carefully consider how to share data effectively while also putting utmost care into protecting human subjects. Making data available for reuse can be costly, particularly given the need to protect privacy. NAHDAP developed archiving procedures for managing and disseminating data requiring protection that evaluate the cost to relative benefit—instituting a process to determine priority and articulating different curation standards to ensure costs are commensurate with the anticipated value of the data. Finally, for shared data to contribute to meaningful public health advances, they must be made available in a way that is useful and understandable, including associated documentation, tools, and technical assistance to ensure appropriate reuse.

Increasing use can also increase privacy concerns. NAHDAP has addressed those effectively by incorporating and sometimes extending the practices and policies of ICPSR, a social and behavioral science archive where it is housed. ICPSR is an international consortium of more than 750 academic institutions and research organizations providing

leadership and training in data access, curation, and analysis methods for the research community. ICPSR maintains a data archive of more than 250,000 files from nearly 18,000 studies and hosts 21 specialized collections of data in education, aging, criminal justice, and other fields than addiction.

2. Background and Important Considerations

As the lead federal agency supporting scientific research on drug use and its consequences, NIDA has the mission to advance science on the causes and consequences of drug use and addiction and to apply that knowledge to improve individual and public health. Data related to drug use are highly sensitive due to the stigmatization of substance misuse, the illegal nature of some of the behaviors reported, and the potentially damaging repercussions if data are not properly protected. The archive also includes research on another sensitive topic, HIV, which in some instances is closely related to substance use. While all human subjects' data need to be handled with care to protect the identity of the persons involved in the research, the highly sensitive nature of data in research supported by NIDA requires the utmost care in evaluating risk and ensuring protection. Due to this concern, as NIDA and ICPSR developed the data archive, we had a keen interest in ensuring that we set the highest standards for protecting the privacy and minimizing data disclosure risk

Beyond primary concerns around reducing disclosure risk, as stewards of federal dollars, it was important to consider the costs associated with archiving data and developing approaches that would best prioritize reuse. Not all data have equal potential for reuse or supporting scientific advances. This translated into carefully considering what kind of data should be enhanced and to what extent. In addition, we needed to develop criteria we would use to appraise the potential reuse value of the data (e.g., breadth and depth of measures, methodological rigor). As part of this prioritization work, it became apparent that retaining a higher level of detail around the demographic and geographic characteristics of the study sample was desirable for reuse potential. This required the archive program to expand its methods for managing secure access to confidential and sensitive data so they, too, could be reused.

The final compponent of the goal of increasing data sharing was ensuring and increasing data reuse. From the beginning, the project implemented a data curation approach to ensure the quality and usability of the data by the research community. As archives require users to meet the rigorous demands for accessing sensitive and confidential data, ensuring the data can be used is doubly important. The NAHDAP project leveraged emerging metadata standards, practices, and tools for improving data reuse. In addition to data curation, NIDA provided funding for technical assistance for both data depositors and users to maximize the impact of the data and resources, allowing users to navigate better the systems we put in place for both depositing and disseminating data.

The parameters that guided us during the development of the archive are roughly grouped into three areas: striking a balance between high value placed on protecting human subjects' sensitive data and making data available; careful consideration of the costs of archiving relative to the value of the data, including costs to ensure privacy; and ensuring data were available in a way that facilitates the greatest use. We have developed this perspective to share some of our lessons learned over the last 13 years. It is intended to be useful to organizations managing confidential human subject data and to any investigator

seeking an ultimate home for their data by raising some points to consider when deciding how to archive and manage human subjects' data for maximal reuse.

3. How sensitive data are protected and shared

Any enterprise focused on the sharing of data must include considerations of minimizing risk to human subjects. When data are highly sensitive, these considerations rise to a new level. The following section outlines the major areas in which we worked to ensure data were protected while also made as widely available as possible.

3.1. **Levels of Access.** One important step in protecting privacy is to consider what level of protection is needed to disseminate the data. NAHDAP releases data through Public Use Files (PUFs), Restricted Use Files (RUF)s, and a Virtual Data Enclave (VDE). Often, a dataset will have both a PUF and a RUF. When data are sensitive and privacy concerns are paramount, managing access to a RUF involves, at a minimum, requiring from users a data use agreement with a data security plan. When data are deemed most sensitive, it is possible to impose additional social and technical controls to keep the data safe such as a virtual (or physical) data enclave. For example, data from the Population Assessment of Tobacco and Health study (PATH), a nationally representative longitudinal study of tobacco use behaviors, attitudes, and health outcomes in US youth and adults, are disseminated through both PUFs and RUFs. While the data in both versions are very similar, the PUF dataset version does not have the same detail as the variables in the RUF dataset, with variables either coarsened or removed in the PUF. In addition, several types of information, including biomarkers, state identifiers, and UPC codes, are not available in the PUF. To ensure the appropriate use of the PATH study RUFs, data users apply to use specific datasets, indicating their research credentials and their intended research questions. Once granted access to the data, researchers must perform all analyses in a Virtual Data Enclave (VDE), which is a secure environment managed by the archive that allows access to the RUFs along with software for data analyses and output generation. Any output must be reviewed by the archive for potential disclosure risk, and only approved output is permitted to be used in oral or written presentations.

3.2. **Data Governance Framework.** Another important step in protecting human subjects is considering the Data Governance Framework for the archive to manage RUFs safely. Ethical data sharing ensures that human subject protections provided during data collection are clear around the topic of data sharing and are followed by the archive disseminating the data. To ensure continuity in the protection of human subjects' information when data are deposited in NAHDAP, NAHDAP staff provide technical assistance to the original research team by reviewing the original consent form, advising on whether any additional consents are needed to share data, and managing data use agreements with the institution where the data were collected and with the institutions of all third-party users. For researchers who have yet to collect data, technical assistance is provided to support writing informed consent to address data sharing and planning different aspects of the data collection as another way the archive can be helpful to ensure continuity.

3.3. **Data Use Agreement for Deposit.** NAHDAP uses two types of data use agreements (DUAs): one for the *deposit* of restricted-use data into NAHDAP and one for the *use* of restricted-use data. DUAs used to deposit a copy of restricted data into an archive are negotiated and signed between the depositor and the archive, resulting in an agreement that the archive will follow when managing the data and providing access to the data by future third-party users. One of the goals NAHDAP pursued was to design a universal deposit agreement for restricted-use data (referred to as the Restricted-Use Data Deposit and Dissemination Agreement or RUDDDA) that would cover the concerns of most researchers and their institutions when depositing data. However, sometimes it is necessary to change or add to the terms of the RUDDDA to be consistent with the language from the original data collections' informed consent and other obligations justified by the depositing institution. For example, some depositing institutions require notification if a future user violates the terms of the signed data use agreement, while other depositing institutions may be silent on this issue, allowing the archive latitude in tracking and managing suspected violations.

3.4. **Data Use Agreement for Users.** The DUA with the depositing institution must then be mapped to the second type of DUA to be signed by the end data users and their institutions. DUAs for users outline the responsibilities of researchers using the data and describe the limitations of using existing data. Again, it is ideal if these DUAs can be standardized across the archive, but in some cases, privacy or other concerns will demand some variation in DUAs within an archive. NAHDAP's DUA for users of restricted-use data requires the users to have and follow data security plans outlining expected and acceptable physical and social security controls that must be in place when using data. RUFs, whether the data may be securely sent to the user or must be accessed within NAHDAP's VDE, both require a data security plan attached to the DUA. For example, all RUF users must use the data within a private space, but whether they can provide the hardware to store and access the data (as is the case for some NAHDAP RUFs) or if they are required to connect to NAHDAP's VDE to access the data (as in the case of PATH study RUFs, for example) is described in the data security plan.

3.5. **Restricted Data Training.** While the DUA and the data security plan set the parameters for using the RUF, it is important to ensure that the agreements are understood by users (Green and Ritchie, 2022). NAHDAP developed training to ensure the acceptable use of data following the terms of the DUA and data security plans. For example, before gaining access to a RUF in the VDE, users must complete an initial training module covering: avoiding violations (e.g., manually copying data down, taking photographs, sharing login information with others) and understanding the consequences of violating the agreements (e.g., terminating access to the RUF, reporting violations to NIDA). Refresher training for users is required every two years.

3.6. **Disclosure Risk Review to Determine Mode of Access.** As data are submitted for deposit, a critical step by the archive for ensuring privacy concerns are addressed is the data disclosure risk review (DRR). This review is conducted to evaluate the risk for reidentification of human subjects and to identify what measures must be taken to reduce this risk. NAHDAP conducts a DRR of all deposited data to ensure that direct identifiers are removed and indirect identifiers are mitigated, and that access to a restricted version

of the data is managed securely. The review is calibrated to the modality of how the data will be accessed (e.g., downloadable versus made available to authorized users with a data security plan in place), the sensitivity of the topics in the survey (e.g., stigmatizing or illegal behaviors), the methodology of the study (e.g., qualitative data have unique security risks), and the potential for linkage with other datasets. Currently, there are no specific practices in place for fully evaluating risk when data are combined. However, the field must be aware that linkage can increase risk, and should work to ensure that consideration of this risk informs how data are protected as the sharing of data increases. As an additional precaution, an ICPSR Disclosure Review Board (DRB) provides guidance and review of disclosure risk concerns that go beyond standard practices used by NAHDAP. The DRB reviews data disclosure policies and practices, acts as an appeals body for disclosure decisions, and advises on new disclosure protection technologies or processes. The DRB strives to balance data protection and the analytic utility of data sets, including protecting individuals and groups.

3.7. **Disclosure Risk Review to Authorize Output.** A way to further protect the privacy and reduce risk to participants is to set rules that apply to using and reporting the results from analysis of the RUF outside the environment described in the data security plan. These rules ensure that there are no unique individuals (or groups) in descriptive results, which could result in re-identification of the human subjects in the study. The most common rule applied to output to be used outside the controlled analytic environment is cell suppression (Cox, 1980) where data are suppressed that fall below a threshold. While there is no absolute standard in the field, minimum cell size thresholds (the minimum number of observations in each cell) are most frequently set at 10, though 5 and 20 are also often used (Ritchie, 2022). The Centers for Medicare and Medicaid Services (CMS) sets the minimum cell threshold at 11 or more for limited data from Medicare and Medicaid records (Center for Medicare and Medicaid Services, 2020). The National Center for Health Statistics recommends recoding variables with counts of five and below (National Center for Health Statisics, 2019). The National Center for Education Statistics requires a minimum cell count of three when reporting results from restricted data (Center for Education Statistics, 2011) and suppresses cell counts below 30 in its remote access system. In the case of NAHDAP, varying output rules (e.g., cell size) and modes of access are set according to the sensitivity of the RUF. RUF users (for select RUF data) who are approved to follow a data security plan managed locally must self-vet the output of their analyses before presenting and publishing the results. For RUF data that must be accessed through the VDE, NAHDAP's research staff reviews and authorizes output before users can use the output. It is important to consider the sensitivity of the data and set a cell size limit that is appropriate for the data, while also balancing those limits with the utility of the data. In addition, the field must understand the dangers inherent in publishing data with low cell numbers. Journal editors, reviewers, and manuscript managers should ensure that small cell sizes are not published.

3.8. **Incident Response Protocol for Potential Data Security Violations.** Finally, a data archive must develop a plan to enforce data use agreements if they discover that data or output may have been used in violation of the agreement. Without a way to enforce agreements, individuals violating terms of use have no repercussions if they violate the ethical obligations or terms of use defined in the DUA. There is a general expectation in the scientific community that most researchers are trained in human subject ethics and would

not knowingly violate a data use agreement. However, our experiences providing access to RUF data have demonstrated that RUF users may unintentionally violate the terms specified in the DUA. It can be difficult to discover these kinds of protocol violations, especially if they happen outside a VDE. This may be an area requiring additional research to understand the magnitude of the problem and to develop effective solutions to mitigate violations. We learned that even with the emphasis on user support, communication with users, training, and VDE monitoring, there were many more instances of misuse than anticipated. This points to an important sociotechnical control for enforcement which is ensuring that staff managing these systems are available to answer questions and guide data users through compliance with their data use agreements. Most DUA violations discovered result, at least partly, from misunderstanding the terms of the agreement (e.g., sharing a password with a new research assistant) or simple errors (e.g., failing to notify NAHDAP when moving to a new institution). When violations are discovered, NAHDAP requires users to refresh their knowledge through one-on-one consultation with NAHDAP staff or retaking the required VDE training module before resuming access to and use the data.

An important part of NAHDAP's history and experience arose from uncovering intentional violations of the DUA, where more than one user manually recorded output from the VDE despite knowing that this was a violation. This led NAHDAP to develop additional policies laying out clear consequences to respond to this direct and seemingly intentional violation of the data use agreement. In developing these additional policies, we queried other archives to understand what policies were in place and found that there were no clear standards for enforcing DUAs or for the actions to take in the instance of a violation. In cases where archives did have actions, there was no clear path to implementation, nor could they provide examples where they had enforced an action. NAHDAP developed an enforcement protocol to restrict the users who were in violation from accessing any RUF data within NAHDAP archive. Beyond restricting individual users, given the seriousness of the violations, NAHDAP also leveraged the ability to ban all investigators from institutions where the violations occurred. We consider it crucial that the institutions demonstrate an understanding of the seriousness of the ethical violation and indicate that they would comply fully with the terms of the DUA. Once the institutions came back into compliance, users in good standing at those institutions were again granted full access to ICPSR data. As NAHDAP and the broader ICPSR organization house a range of sensitive data used by many researchers, this enforcement policy was sufficiently limiting and impactful that the institutions came into compliance. This experience points to the need to consider the enforcement of DUAs carefully and to develop clear policies to handle violations. These policies must be implementable and detail exactly how implementation will work.

4. Managing Costs for Data Sharing

Resources are needed to support the work necessary to ensure privacy protections and enhance the utility of shared data. When data are released in a format that is not usable, there will be significant challenges in attempts to reuse the data for additional analyses. NAHDAP has set forth a goal that the most potentially useful data will be documented thoroughly (e.g., study methods well described, various elements of the data are labeled clearly, and detailed codebooks produced), standardized for consistency and ease of use, human subjects' concerns adequately addressed and enhanced for findability and usability. Data archives and data producers determine which entity will take on which elements of

this data preparation work to ensure privacy and utility. NIDA and NAHDAP have had to carefully consider the costs for the archive that are involved in curating each dataset, developing systems to assess value, decreasing costs where possible, and thus balancing costs of data sharing against the anticipated demand for and usefulness of the data.

4.1. **Selecting Data for NAHDAP.** One of the ways that we reduce costs is by working to archive the most valuable datasets. This involves assigning a level of reuse value to the dataset. To determine whether we would use limited project resources to support archiving a dataset, we defined what types of studies would be valued most highly relative to the subject matter goals of the archive. We developed criteria by which we could evaluate data and applied these criteria to research projects currently or recently funded by NIDA (as well as any data offered to NAHDAP without invitation). Criteria included longitudinal data or any program that NIDA had invested in long-term; whether the participants were unique or pertain to a group under-represented in research; and inclusion of relevant populations/measures (e.g., persons engaged in substance use/misuse). We used these criteria to inform consideration of value but also used a subjective assessment of other more unique aspects of the project (e.g., emphasis on an understudied population, fit with other data at NAHDAP) to inform our ultimate decision on which data NAHDAP would invest in archiving.

4.2. **Selecting a Curation Level.** In addition to the data selection process, a related step to ensure that we are investing resources efficiently was considering the amount of work needed to curate a dataset to enhance its utility maximally. Based on this, we assign each deposited dataset a curation level that reflects the amount of review, checking, and detailed work on the files that will be performed by the repository. The lowest level, Level 1, is selected when data need little additional curation. Levels 2 and 3 include a more nuanced review and mitigation of disclosure risks, additional data checks, processing steps, and detailed work. Beyond the completeness and quality of the deposited data and documentation, the expected reuse value of the data was considered when assigning a curation level. We assign resources to curate data if the balance between value and cost seems reasonable. For example, we would only invest in the highest level of curation if we had determined that the data were likely highly valuable or were not usable in the form they were deposited to the archive. In cases where a high level of curation is needed, the archive often still takes in these data but would reduce cost by releasing only a public use file (i.e., removing altogether potentially disclosive variables such as open-ended text fields) without the additional work of reviewing and designing other ways to mitigate disclosure risk. In this way, a balance is struck between value and cost. As all data will not be equally valuable for re-use, it may be important to develop review standards for what data should be archived and at what level so that resources are not expended on projects with less long-term value. Pilot projects and other small projects might fall into this category.

4.3. **Technical Assistance during Data Management Planning.** Researchers developing proposals, data management and sharing plans, and designing the collection of data have the opportunity to build resources into their projects and give attention to producing high quality data and documentation, so that less remediation is required when sharing. NAHDAP's team frequently advises, shares resources, and provides training to researchers

at these earlier stages of the data life cycle when the investment in data management stands to have the greatest lasting benefit to the team themselves as well as future data users. With this early intervention, the subsequent cost of depositing and curating data may significantly decrease.

5. Making Data Accessible and Useful

This final section details how NAHDAP invests in the discoverability and reuse of RUF and PUF data in the archive. Although this may reflect a smaller part of the program investment, it means the systems and processes we have implemented through NAHDAP get used.

5.1. **NIDA's Investment.** Simultaneously with establishing NAHDAP, NIDA also published a Funding Opportunity Announcement (FOA), "Accelerating the Pace of Drug Abuse Research Using Existing Data," to support investigators in submitting research applications using archived datasets. The FOA solicits applications that are "proposing innovative analysis of existing social science, behavioral, administrative, and neuroimaging data to study the etiology and epidemiology of substance using behaviors (defined as alcohol, tobacco, prescription drugs, and other substances) and related disorders, prevention of substance use and HIV, and health service utilization. This FOA was renewed in January 2022, and remains open. This FOA highlights NIH's support for data sharing and broadening the use of data by others than the original investigators.

5.2. **Improving the Usability of Data.** NAHDAP's curation processes ensure that data can be used now and into the future. Trained, expert curation staff perform quality checks on the data to identify codes that seem out of line or do not match the documentation, add question text and other variable-level information as needed, create metadata records describing the studies for better discoverability, review the data for potential disclosure risks in terms of sensitive or personally identifying information and work to mitigate such risks. Multiple versions of each dataset are created: an ASCII comma-separated file for preservation and files in the major statistical packages (R, SAS, SPSS, and Stata), and and files online analysis. NAHDAP works with depositing research teams to ensure all information in the data package is correct and complete.

5.3. **Enabling Findability and Searching.** One limitation to using existing data for research is that datasets are often not housed in the same physical or virtual location, requiring investigators to download files from multiple organizations and then merge them. Similarly, when there is no persistent access to data, this can lead to numerous analytic issues, such as not having access to verification of published results. Archiving multiple datasets within the same infrastructure facilitates not only locating specific variables of interest but also combining these datasets. NAHDAP's search and comparison tools are useful for finding compatible variables of interest within and across studies and reviewing methods and other pertinent details of the data collection. In addition, co-locating datasets allows for the centralization of other resources. For example, NAHDAP's VDE provides all authorized users access to various analytic software packages and other tools to facilitate the presentation of findings from data analyses.

5.4. **Increasing Impact of Data Sharing.** In addition to realizing efficiencies for both data depositors and data users, centralized archiving also allows for more efficient and effective assessment of use of the data and other metrics. For example, NAHDAP tracks how often public-use datasets and related documentation are downloaded by unique visitors to the website. In addition. NAHDAP provides recommended citations for the datasets and related resources, which makes tracking publications more straightforward. In addition, the NAHDAP librarian can create accurate and standardized search protocols to understand how those datasets are used. For example, NAHDAP includes citations of a variety of data-related literature, including journal articles, magazine articles, newspaper articles, conference presentations, books, book sections, theses, dissertations, web pages, and other information objects. The criteria for inclusion in this bibliography include analysis of data archived by NAHDAP, and a discussion, critique, or extension of another's data analysis related to those archived data.

6. Conclusion

NAHDAP benefits by being loccated at ICPSR, which has a robust infrastructure for providing data archiving services, along with evolving options for managing access to restricted data requiring that human subject protections remain in place. NAHDAP also benefited from the vision and input of a funding agency that understood the complexities and cost of sharing more sensitive data over 13 years ago. Through this joint effort of ICPSR and NIDA, the archive has grown in size and use, providing access to data from hundreds of studies. The project cultivated data for the archive that required the development of legal agreements and a range of targeted archival practices designed to assure the depositing organizations and investigators that their data would be managed and reused ethically by the archive and any future users. While these solutions necessitated investment, NIDA and NAHDAP's team designed the most cost-effective ways to accomplish these goals as well as ensure maximal discovery and use of valuable data for the field.

As the new NIH Data Management and Sharing Policy is implemented, it will be important to consider balancing the costs of data archiving with the value and potential reuse of data. Cost can be reduced if investigators think about depositing their data and take steps along the way to ready the data for easy deposit into an archive. However, in instances where archiving costs could be quite high, it will be important to assess accurately the value of the data and decide on a level of effort for archiving that balances cost and value.

Separately, it is critical in any data sharing venture to take strict measures to ensure the public trust in participating in health research. Employing strategies that maximize protection of human subjects will be critical. Cell size limits are one step that can be taken to protect data, but this may be insufficient. This is especially true when one considers the potential of combining data with administrative data, which could lead to a higher risk of participant re-identification. Re-identification risk analysis should be conducted, though it may be challenging to fully assess risk when combined with potential, unknown additional datasets. The lessons learned over the last 13 years have resulted in some important insights that both data depositors and data archives will consider carefully as data sharing expands.

References

Centers for Medicare and Medicaid Services (2020). CMS Cell Suppression Policy. https://www.hhs.gov/guidance/document/cms-cell-suppression-policy

Cox, L. H. (1980). Suppression methodology and statistical disclosure control. *Journal of the American Statistical Association*, 75(370): 377-385. https://doi.org/10.1080/01621459.1980.10477481

Federer, L. M., Lu, L., Joubert, D. J., Welsh, J., and Brandys, B. (2015). Biomedical data sharing and reuse: Attitudes and practices of clinical and scientific research staff. *PLOS ONE*, 10(6):e0129506. https://doi.org/10.1371/journal.pone.0129506

Green, E., and Ritchie, F. (2022). Training analysts in the management of confidential data. *Data Research, Access, and Governance Network Working Paper Series*, Number 2022/22. (https://uwe-repository.worktribe.com/output/9258065)

Lohr, S. L., and Raghunathan, T. E. (2017). Combining survey data with other data sources. *Statistical Science*, 32(2):293-312. https://doi.org/10.1214/16-STS584

National Center for Education Statistics (2011). *Restricted-Use Data Procedures Manual*. U.S. Department of Education, Institute of Education Sciences. https://nces.ed.gov/pubs96/96860rev.pdf

National Center for Health Statistics Research Data Center (2019). Disclosure manual: preventing disclosure: rules for researchers. https://www.cdc.gov/rdc/data/b4/Disclosure-Manual-v2.3.pdf

National Institutes of Health, Office of the Director (2020). Final NIH Policy for Data Management and Sharing. NOT-OD-21-013 https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html

Ritchie, F. (2022). Ten is the safest number that there's ever been. *Transactions on Data Privacy*, 15(2):109-140. https://uwe-repository.worktribe.com/output/9853172