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

NAHDAP-ICPSR

Submitted Nov 2022
Published Nov 2023

BEYOND LEGAL FRAMEWORKS AND SECURITY CONTROLS FOR ACCESSING CONFIDENTIAL SURVEY DATA IN THE UNITED STATES: ENGAGING DATA USERS IN DATA PROTECTION

AMY M. PIENTA[†], JOY BOHYUN JANG[†], AND MARGARET C. LEVENSTEIN[†]

[†] ICPSR, University of Michigan

Abstract. With growing demand for data reuse and open data within the scientific ecosystem, protecting the confidentialy of survey data and privacy of data subjects is increasingly important. Doing so requires more than legal procedures and technological controls; it requires social and behavioral intervention. In this research note, we delineate the disclosure risks of various types of survey data (e.g., longitudinal data, social network data, sensitive information, biomarkers, and geographic data), the current motivation for data reuse, and challenges to data protection. Despite rigorous efforts to protect data, there are still threats to protection of confidentiality in microdata. Unintentional data breaches, protocol violations, and data misuse are observed even in well-established restricted data access systems, indicating that the systems may all rely heavily on trust. Creating and maintaining that trust is critical to secure data access. We suggest four ways of building trust; *User-Centered Design Practices; Promoting Trust for Protecting Confidential Data; General Training in Research Ethics; and Specific Training in Data Security Protocols,* with an example of a new project '*Researcher Passport*' by the Inter-university Consortium for Political and Social Research. Continuous user-focused improvements in restricted data access systems are necessary so that we promote a culture of trust among the research and data user community, train both in the general topic of responsible research and in the specific requirements of these systems, and offer systematic and holistic solutions.

1. Introduction

Protecting the confidentiality of data supplied by survey respondents while advancing data reuse and open science goals is a space navigated by federal sponsors, regulatory agencies, statistical agencies, research organizations, data repositories, and others. In this space, legal agreements, data security plans, technological controls, and other means safeguard respondent identities and information contained within survey data while advancing science and knowledge (Abowd, Schmutte and Vilhuber, 2021; De Wolf, 2003; Ritchie, 2017).

Key words and phrases: training, confidentiality, data governance.

Corresponding author: A. Pienta, apienta@umich.edu. This work was partially supported by The National Institute on Drug Abuse (#75N95019C00017), the Alfred P. Sloan Foundation, and the National Science Foundation (#1839868).

www.journalprivacyconfidentiality.org
DOI:10.29012/jpc.842

However, given that restricted data access is complex and often hard to navigate, and well-intentioned scholars face competing demands on their time and incentives for speed in publishing, more than legal and pragmatic solutions may be needed to protect data. This paper discusses the changing context of accessing confidential survey microdata, and offers recommendations for improving user compliance with these systems, in order to safeguard survey microdata better while maximizing research potential.

2. Changing Confidentiality and Sensitivity of Survey Microdata

Survey microdata are generated from structured, closed-ended and open-ended questions measuring experiences, behaviors, attitudes, and attributes from a sample representing an underlying population of interest. An underpinning of social science use of surveys is that identifying information about the survey respondents can be omitted or disguised so that the data can be analyzed without disclosing the identities of survey participants. However, even survey data without direct identifiers are still at risk of disclosure because of advances in research—for example, unique combinations of personal attributes, detailed geography, and increasing capability of linkage to external data (Solomon, et al., 2012).

We discuss five particular issues. First, there is increasing reliance on longitudinal surveys (relative to cross-sectional surveys) to enable researchers to address causality and change over time. When survey respondents provide multiple observations over a period of study, the resulting data contain more detailed information about respondents and their unique combinations of experiences and attributes, introducing greater disclosure risk (Duncan and Stokes, 2004). Longitudinal data cannot always be fully de-identified without undermining their fitness for use, so they often require additional safeguards.

Second, surveys administered to various types of social networks, including dyads, families, households, peer groups, and classrooms, contain more information about the context surrounding an individual; this increases disclosure risk. Particular risks arise when members of the network are unaware of the participation of others in the study. Many large-scale surveys include this design element, such as interviewing (1) a spouse or partner, as in the Health and Retirement Survey (Servais, 2010); (2) adolescents and their parents as, in the National Survey of Adolescent Health (Harris, 2013) and the Population Assessment of Tobacco and Health Study; and (3) families and households over time as in the Panel Study of Income Dynamics (PSID Main Interview User Manual: Release 2019). Standard anonymization techniques may not protect against re-identification when data are collected from related individuals in social networks (Hay, et al., 2007). Such survey methods necessitate implementing controlled access methods to prevent data disclosure.

Third, as research has crossed disciplines, biomarkers, including genetic information, are sometimes collected as part of surveys. Adding biomarker data to a survey introduces different levels of harm that might result from re-identifying the data. For example, re-identification might make detrimental genetic information available to insurance companies or employers (Hansson, et al., 2016). Although most participants want their information to be used and shared with other researchers to provide more benefit, availability of biomarker information in surveys further underscores the importance of ensuring that data are protected and used securely. This is necessary in order to continue the necessary public trust for participating in these kinds of studies.

Fourth, detailed geographic information increases analytic value by facilitating linkage to contextual data and analysis with new data visualization tools. However, detailed geographic

information also increases the risk of re-identification. Where one lives, goes to school and work, seeks health care, as well as social and recreational activities, all condition individual attitudes, behaviors, and experiences. Collecting and retaining geographic location (e.g., state, county, city, and smaller areas of geography such as Census block and tract, address, or latitude/longitude) allow analysis and interrogation of how places and groups matter.

Finally, the unintentional risk of (or even having a plan to facilitate) future data linkage underlies many of the issues already discussed. Combining survey microdata with other data sources increases risk due to increasingly powerful search utilities (Lan, et al., 2011), advancing record linkage methods (Domingo-Ferrer and Torra, 2003), and more publicly available information.

Thus, many traditional means of data de-identification may be unable to protect respondent confidentiality adequately. Controlled access methods have gained traction to help reduce the risk of an intruder disclosing the data in the future.

3. Evolving Requirements of Open Science

Alongside the significant and growing challenges in protecting confidentiality in survey data, there is an increasing demand within the scientific ecosystem for access, transparency, and replicability of data and methods (Lupia, 2020). Government agencies, such as the National Institutes of Health (NIH) and the National Science Foundation (NSF) in the United States, and many others within and outside governments across the world, wish to ensure broad access to data originating from their funding, so that the data can be used for purposes well beyond what had been proposed by the original study team (e.g., OECD Open Science (https://www.oecd.org/sti/inno/open-science.htm); UNESCO Open Science (https://www.unesco.org/en/natural-sciences/open-science). Specifically, the expectation is for data to be made available publicly in support of transparency and reproducibility, to generate new findings, and to inform the development of new measures and methods for training and education. Equally important is that data be harmonized with, integrated with orlinked to other available data, such as administrative data and data from other disciplines (Lohr and Raghunathan, 2017). There is also growing credit for public data sharing as we see dedicated repositories assign persistent identifiers for data, formalizing credit for authoring data (Bierer, et al., 2017) and allowing better tracking of use and impact of data.

In the United States, there has been increasing recognition of the tremendous potential of existing research data to be used by secondary analysts to address new research questions, provide policy-relevant findings, and, when data are combined, yield new information to guide policy and practice. A decade ago, the Office of Science and Technology Policy, within the Executive Office of the President, issued an executive memorandum entitled *Increasing Access to the Results of Federally Funded Scientific Research* (OSTP, 2013). In it, federal agencies spending more than 100 million dollars annually in research and development are required to draft plans to increase public access to research results, including public access to the original data. The alignment of academic research communities, sponsors, and data repositories behind the OSTP memo and its legislation goals is critical to achieving data sharing and access (Ember and Hanisch, 2013). The NIH recently released the updated Data Management and Sharing Policy, which requires the submission of Data Management and Sharing Plans for all research funded or conducted by NIH that results in the generation of scientific data (https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html). Other

funding agencies in the US also require the sharing of scientific data (e.g., NSF data sharing policy: (https://www.nsf.gov/bfa/dias/policy/dmp.jsp); US Congress Foundations for Evidence-Based Policymaking Act of 2018 (https://www.nsf.gov/bfa/dias/policy/dmp.jsp)). There are important ethical reasons for open science and data sharing to promote transparency and integrity, perhaps now more than ever (Lupia, 2020).

Evidence suggests that there is a return to be gained from data sharing. Data sharing increases the impact of the original datasets and enables downstream impact through data reuse. Papers with archived datasets receive more citations (Piwowar and Vision, 2013; Piwowar, et al., 2007; Pienta, et al., 2010), increasing the impact of those papers relative to those without shared data. Funders and research participants have the right to expect that their data will be used whenever possible to build knowledge (Walport and Brest, 2011), and recent studies show participants, especially, want their data to be used for research provided that their privacy is maintained (Damschroder, et al., 2007; Fiesler and Proferes, 2018).

Despite the high value of access to secondary data and reuse, it is important to underscore socio-technical challenges that scientific communities must resolve. Research is an increasingly data-intensive practice, where we see more scientists generating or relying on data in their day-to-day pursuit of scholarship and knowledge (Borgman, 2016). Secure data access is paramount and challenging to achieve in light of

(1) Researcher objections to obligations to share data (Oushy, et al., 2015);
(2) Institutional Review Boards (IRBs) that operate under their local understanding of data protection (Goldenberg, et al., 2015);
(3) Increased need for adequate informed consent regarding data sharing (Hornstein, et al., 2015; Macilotti, 2013);
(4) The complexity of various approaches to protect data;
(5) Changing demographics of the populations being studied and their views toward data privacy (Gfroerer, et al., 2003; Tennant, et al., 2015).

These all suggests that any data-sharing model must continually evaluate how best to safeguard human subjects' data.

4. Modes of Data Protection

Survey microdata often undergo some form of statistical disclosure limitation to protect the privacy of survey respondents and confidentiality of their data (Shlomo and Skinner, 2010). In some instances, the process limits the value to analysts. Survey microdata requiring disclosure risk protection are often disseminated using a combination of legal and technical measures complementing public-use sets. Restricted data access is typically managed by the data provider or a third party, such as a data repository. Data are available to the research community only after what can be a lengthy and complicated application process. Restricted data access models rely on qualifying researchers wishing to use data (Grayson, et al., 2019). Tyler (2020) finds that data repositories handling restricted data evaluate potential users of confidential data along four common dimensions, including identity (when anonymous data access is not allowed), training (such as holding an advanced degree), reputation (institutional affiliation), and aspects of the project. Across various dissemination models, processes need to be more consistent, not only in what they require of researchers, but also in how they define other things, such as modalities of access and what constitutes responsible and trusted users (Tyler, 2020). Thus, restricted data access systems may lead to error and misunderstanding.

Restricted data access systems often require applicants to describe the computing environments that their institutions provide. In some cases, prospective data users are limited to a specific computing environment that the data provider hosts (e.g., a virtual data enclave), which may be limited in available software or computational capabilities. A required data security plan specifies the rules, processes, and location for accessing and analyzing data (e.g., required IRB review, researcher training, and researcher and institutional safeguards). The plan aims to ensure the security of the data if they leave the providers' facilities. And yet, adhering to this plan requires users' and organizational understanding. It is critical, therefore, that data stewardship organizations such as archives have tools to enforce these plans when they are implemented.

5. Threats to our Systems for Accessing Confidential Data

Despite the availability of sophisticated systems that provide legal and technological protections for confidential data (e.g., physical and virtual enclaves such as the Federal Statistical Research Data Centers or Inter-university Consortium for Political and Social Research (ICPSR)'s Virtual Data Enclave), there are still threats that need to be mitigated. These systems generally protect in two ways: limiting what can be brought in and limiting what can be taken out. Limiting what can be brought in, by prohibiting access to the Internet or local computing drives, for example, makes it much more difficult to re-identify data by combining them with other data. Limiting what can be taken out allows for third-party disclosure review to ensure that anything removed from the secure computing environment meets the standards for safety (e.g., what was promised to study participants or is required by regulations).

In addition to these computing environments' technological protections, data stewards generally use legal agreements to protect data confidentiality. Researchers sometimes must sign agreements to protect confidentiality for life, with potential criminal liability (fines or imprisonment). In other cases, researchers and their host institutions sign data use agreements that expressly commit the researchers to protecting confidentiality, but may or may not have explicit consequences associated with violating the agreement. Regardless, vulnerabilities remain even with these legal and technical protections in place.

First, data breaches, though uncommon, may occur if restricted data are removed from an authorized system. For example, an intruder may maliciously penetrate an authorized system and remove confidential data. It is also possible that an authorized user may intentionally or unintentionally write down information or use a cell phone to capture a snapshot of data containing identifiers.

The second, more common, kind of confidential data violation is a protocol violation that stems from a lack of adherence to data security plans, including rules to ensure data are safeguarded at all times and used only as authorized. Using data on a networked computer, sharing a password with another individual who is not authorized, failing to notify the data provider when changing affiliations, and other behaviors are examples of breaking the procedural requirements outlined in a data use agreement, and thereby compromising data confidentiality.

Third, misuse may arise due to the complexity of obligations one takes on when managing restricted data, which creates challenges that put the data and the individuals' information within it at risk for disclosure. It is known that users develop insecure workarounds in complex cybersecurity systems (Sinclair and Smith, 2010; Blythe, et al., 2013). System

complexity may hinder science and leave research data vulnerable to unauthorized use. Many have argued that some users will always find, intentionally or not, ways to circumvent data security. This means that established restricted-data access systems all rely heavily on the trust placed in users. Developing additional complex, technical solutions may not be the answer to increasing security of confidential data, but instead, we should strive to grow a culture of trust.

6. Keeping Confidential Data Not Just Safe but Even Safer

Increasing user acceptance, understanding, and trust in restricted data access will likely come from better-designed systems where users understand their expected behaviors in those systems and their role in ensuring the integrity of those systems. Restricted data access systems are often managed by data repositories such as ICPSR, where the data repository accepts and manages the original ethical obligations of informed consent. ICPSR, like others managing access to restricted-use data, safeguards data using a virtual data archive (VDE) in which microdata remains within the enclave until authorized for use outside the VDE. While there have been rare instances in which restricted-use data from ICPSR have been breached, minor violations, such as protocol violations, are more frequent and deserve attention. We offer suggestions to improve how users interact with restricted data access systems to protect confidential data better.

6.1. **User-Centered Design Practices.** The systems to access confidential data have already been described as complex. A better understanding of the workflows of researchers accessing confidential data, how they work, and with whom they work or collaborate should inform future design. Involving end users in developing these complex systems would increase acceptance (Karlsson and Hedstrom, 2014). For example, there has been a rise in collaboration among scientists, even across disciplines (Wagner, et al., 2018), and systems for accessing confidential microdata must accommodate project teams, including those less familiar with data privacy and the obligations and protections built into those systems. Addressing the needs of a wide range of users might improve the understanding and efficacy of confidential data users and avoid system workarounds that increase the chance of unauthorized disclosure.

6.2. **Promoting Trust for Protecting Confidential Data.** Organizations providing access to confidential data usually have ample sanctions in place, but place less emphasis on ensuring that users know how to protect data and on motivation or a sense of shared responsibility and ethos (Feth, et al. 2017). Researchers who use secondary data face pressures to make steady progress and publish results. This reality might be considered at odds with restricted data access systems, which introduce delays in gaining data access and delay one's ability to use research outputs until authorized. Opportunity exists to encourage users of confidential data to recognize that their behaviors impact the continuation of those systems: accessing confidential data is a privilege requiring their active role in protecting data privacy. In this vein, SYNAPSE, a research collaboration platform at Sage Bionetworks, asks users to sign an oath that asks for recognition as being part of a community, verifies primacy of privacy laws and regulations in guiding behavior, reinforces ethical responsibilities, and requires their adherence to technical and administrative measures. ICPSR's Researcher

Passport (Section 6.5) is another example of recognizing and encouraging trust among data users.

6.3. **General Training in Research Ethics.** Some gaps in general research ethics training affect data privacy. U.S. federal research funding agencies, e.g., NIH and NSF, require those who receive research funding to complete responsible conduct of research training; only the NIH has stated explicitly what that training should include. Data providers may assume that academic researchers have completed training that meets federal requirements. However, the reporting and explicit requirements for researcher training in responsible conduct of research, proper data security practices, and data management vary substantially across data providers and repositories. The relevant training curricula include data privacy and confidentiality, responsible data use, information security, enclave access, disclosure control, and data stewardship.

6.4. **Specific Training in Data Security Protocols.** Beyond general training in responsible research, specific training in data security protocols would increase compliance with requirements. For example, ICPSR requires biannual training of all users of their VDE, which covers expected practices and behaviors that users are expected to manage, such as getting all output reviewed and authorized before use outside the VDE, not sharing passwords, allowing adequate time for review of output, and not writing down results in a notebook.

6.5. **Researcher Passport.** ICPSR's Researcher Passport is a new project that is meant to develop infrastructure to expedite access to restricted data. It allows users to build a reputation as responsible users of confidential data in these complex systems, provides a new dimension for enforcement for the myriad of requirements placed on confidential data users, and provides links to publications which are verified safe outputs of the confidential data access system giving the researchers credit and visibility for their actions (Levenstein, et al., 2018). Much like a passport used for border control, ICPSR's Researcher Passport records verified credentials and accumulated user experience interacting with confidential data. This new technology incorporates many of the above recommendations into its design and rewards users for maintaining their credentials, training, and responsible use of confidential data.

7. Conclusion

Survey microdata are unlikely to become easier to protect from disclosure in the future, especially as methods for collecting survey data test the boundaries of what we can reasonably protect with statistical disclosure techniques. At the same, many organizations have invested in designing systems for providing restricted data access to a managed set of users, and the number of organizations offering such systems is rising. Examining the shortcomings of existing systems suggests that further improvements targeting users of these systems are necessary so that we design with a broad group of users in mind. Doing so promotes a culture of trust among the research and data user community, train both in responsible research and also in the specific requirements of these systems, and offer new solutions such as the Researcher Passport at ICPSR that will address the weaknesses more systematically

and holistically. Advances in statistical disclosure controls and restricted data access are needed to keep pace with the changing threats to survey microdata.

References

Abowd, J. M., Schmutte, I. M., and Vilhuber, L. (2021). Disclosure limitation and confidentiality protection in linked data. *Administrative Records for Survey Methodology*, 25-59. https://doi.org/10.1002/9781119272076.ch2

Bierer, B. E., Crosas, M., and Pierce, H. H. (2017). Data authorship as an incentive to data sharing. *New England Journal of Medicine,* 376:1684-1687. https://doi.org/10.1056/NEJMsb1616595

Blythe, J., Koppel, R., and Smith, S. W. (2013). Circumvention of security: good Users do bad things. *IEEE Security Privacy,* 11:80-83. https://doi.org/10.1109/MSP.2013.110

Borgman, C. (2016). *Big Data, Little Data, No Data: Scholarship in a Networked World,* Cambridge, MA: MIT Press. ISBN: 9780262529914

Damschroder, L. J., Pritts, J. L., Neblo, M. A., Kalarickal, R. J., Creswell, J. W., and Hayward, R. A. (2007). Patients, privacy and trust: patients' willingness to allow researchers to access their medical records. *Social Science & Medicine,* 64:223-235. https://doi.org/10.1016/j.socscimed.2006.08.045

De Wolf, V. A. (2003). Issues in accessing and sharing confidential survey and social science data. *Data Science Journal*, 2:66-74. https://doi.org/10.2481/dsj.2.66

Domingo-Ferrer, J., and Torra, V. (2003). Disclosure risk assessment in statistical microdata protection via advanced record linkage. *Statistics and Computing,* 13:343-354. https://doi.org/10.1023/A:1025666923033

Duncan, G. T., and Stokes, S. L. (2004). Disclosure risk vs. data utility: the RU confidentiality map as applied to topcoding. *Chance*, 17:16-20. https://doi.org/10.1080/09332480.2004.10554908

Ember, C., and Hanisch, R. (2013). Sustaining domain repositories for digital data: A white paper. https://hdl.handle.net/2027.42/136145)

Feth, D., Maier, A., and Polst, S. (2017). A User-Centered Model for Usable Security and Privacy. In *International Conference on Human Aspects of Information Security, Privacy, and Trust.* Springer, Cham, 74-89. https://doi.org/10.1007/978-3-319-58460-7_6

Fiesler, C., and Proferes, N. (2018). Participant perceptions of Twitter research ethics. *Social Media + Society,* 4:1-14. https://doi.org/10.1177/2056305118763366

Gfroerer, J., Penne, M., Pemberton, M., and Folsom, R. (2003). Substance abuse treatment need among older adults in 2020: the impact of the aging baby-boom cohort. *Drug and Alcohol Dependence*, 69:127-135. https://doi.org/10.1016/S0376-8716(02)00307-1

Goldenberg, A. J., Maschke, K. J., Joffe, S., Botkin, J. R., Rothwell, E., Murray, T. H., Anderson, R., Deming, N., Rosenthal, B. F., and Rivera, S. M. (2015). IRB practices

and policies regarding the secondary research use of biospecimens. *BMC Medical Ethics*, 16:1-8. https://doi.org/10.1186/s12910-015-0020-1

Grayson, S., Suver, C., Wilbanks, J., and Doerr, M. (2019). Open Data Sharing in the 21st Century: Sage Bionetworks' Qualified Research Program and Its Application in mHealth Data Release, Available at SSRN 3502410. https://doi.org/10.2139/ssrn.3502410

Hansson, M. G., Lochmüller, H., Riess, O., Schaefer, F., Orth, M., Rubinstein, Y., Molster, C., Dawkins, H., Taruscio, D., Posada, M., and Woods, S. (2016). The risk of re-identification versus the need to identify individuals in rare disease research. *European Journal of Human Genetics*, 24(11):1553-1558. https://doi.org/10.1038/ejhg.2016.52

Harris, K. (2013). *The Add Health Study: Design and Accomplishments*, Chapel Hill: Carolina Population Center, University of North Carolina at Chapel Hill, 1-22. https://doi.org/10.17615/C6TW87

Hay, M., Miklau, G., Jensen, D., Weis, P., and Srivastava, S. (2007). Anonymizing Social Networks, *Computer Science Department Faculty Publication Series*, 180. https://doi.org/10.1201/9781420091502-c15

Hornstein, D., Nakar, S., Weinberger, S., and Greenbaum, D. (2015). More nuanced informed consent is not necessarily better informed consent. *The American Journal of Bioethics*, 15(9):51-53. https://doi.org/10.1080/15265161.2015.1062167

Karlsson, F., and Hedström, K. (2014). End user development and information security culture. In *Human Aspects of Information Security, Privacy, and Trust: Second International Conference, HAS 2014.* Held as Part of HCI International 2014, Heraklion, Crete, Greece, June 22-27, 2014. Proceedings 2 (pp. 246-257). Springer International Publishing. https://doi.org/10.1007/978-3-319-07620-1_22

Lan, C. W., Chen, Y. H., Grandison, T., Huang, A. F., Chung, J. Y., and Tseng, L. F. (2011). A privacy reinforcement approach against de-identified datasets. In *2011 IEEE 8th International Conference on e-Business Engineering* (pp. 370-375). IEEE. https://doi.org/10.1109/ICEBE.2011.25

Levenstein, M. C., Tyler, A. R., and Davidson Bleckman, J. (2018). The researcher passport: Improving data access and confidentiality protection. *ICPSR White Paper Series,* 1. Ann Arbor, MI: University of Michigan Inter-university Consortium for Political and Social Research. https://hdl.handle.net/2027.42/143808)

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

Lupia, A. (2020). Practical and ethical reasons for pursuing a more open science, *PS: Political Science & Politics* 1-4. https://doi.org/10.1017/S1049096520000979

Macilotti, M. (2012). Informed consent and research biobanks: a challenge in three dimensions. In *Comparative Issues in the Governance of Research Biobanks: Property, Privacy, Intellectual Property, and the Role of Technology* (pp. 143-161). Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-33116-9_9

Office of Science and Technology Policy (2013). *Increasing Access to the Results of Federally Funded Scientific Research.* Executive Office of the White House OSTP Memorandum.

https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

Oushy, M. H., Palacios, R., Holden, A. E., Ramirez, A. G., Gallion, K. J., and O'Connell, M. A. (2015). To share or not to share? A survey of biomedical researchers in the US Southwest, an ethnically diverse region. *PloS One*, 10:e0138239. https://doi.org/10.1371/journal.pone.0138239

Pienta, A. M., Alter, G. C., and Lyle, J. A. (2010). The Enduring Value of Social Science Research: The Use and Reuse of Primary Research Data, *ICPSR Working Paper*. http://hdl.handle.net/2027.42/78307

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

Piwowar, H. A., Day, R. S., and Fridsma, D. B. (2007). Sharing detailed research data is associated with increased citation rate. *PloS One* 2:e308. https://doi.org/10.1371/journal.pone.0000308

PSID (2019). *Main Interview User Manual: Release 2019*. Institute for Social Research, University of Michigan, February, 2019. https://psidonline.isr.umich.edu/data/Documentation/UserGuide2017.pdf

Ritchie, F. (2017). The 'Five Safes': a framework for planning, designing and evaluating data access solutions. *Data for Policy*. https://uwe-repository.worktribe.com/index.php/preview/880718/99_Ritchie.pdf

Servais, M. A. (2010). Overview of HRS Public Data Files for Cross-sectional and Longitudinal Analysis. *Ann Arbor, Michigan: Survey Research Center, Institute for Social Research, University of Michigan.* https://hrs.isr.umich.edu/sites/default/files/biblio/OverviewofHRSPublicData.pdf

Shlomo N., and Skinner, C. (2010). Assessing the protection provided by misclassification-based disclosure limitation methods and survey microdata. *Annals of Applied Statistics*, 4(3):1291-1310. https://doi.org/10.1214/09-AOAS317

Sinclair, S., and Smith, S. W. (2010). What's wrong with access control in the real world? *IEEE Security Privacy*, 8:74-77. http://doi.org/10.1109/MSP.2010.139.

Solomon, A. C., Hill, R., Janssen, E., Sanders, S. A., and Heiman, J. R. (2012). Uniqueness and how it impacts privacy in health-related social science datasets. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium* (pp. 523-532). https://doi.org/10.1145/2110363.2110422

Tennant, B., Stellefson, M., Dodd, V., Chaney, B., Chaney, D., Paige, S., and Alber, J. (2015). eHealth literacy and Web 2.0 health information seeking behaviors among baby boomers and older adults. *Journal of Medical Internet Research*, 17:e70. https://doi.org/10.2196/jmir.3992

Tyler, A. (2020). Facilitating access to restricted data. *International Journal of Digital Curation,* 15:1-16. https://doi.org/10.2218/ijdc.v15i1.602

Wagner, C. S., Wagner, C. S., and Graber (2018). *Collaborative Era in Science*. London: Palgrave Macmillan. https://doi.org/10.1007/978-3-319-94986-4

Walport, M., and Brest, P. (2011). Sharing research data to improve public health. *The Lancet,* 377:537-539. https://doi.org/10.1016/S0140-6736(10)62234-9