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

NAHDAP-ICPSR

Submitted Nov 2022
Published Nov 2023

RESTRICTED DATA MANAGEMENT: THE CURRENT PRACTICE AND THE FUTURE

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

[†]ICPSR, University of Michigan

Abstract. Many organizations across the world that manage restricted data have adapted the Five Safes framework (safe data, safe projects, safe people, safe setting, safe output) for their management of restricted and confidential data. While the Five Safes have been well integrated throughout the data life cycle, organizations encounter several challenges with regard to safe data management. In this paper, we review current practices for restricted data management, and discuss challenges and future directions. We focus on data use agreements, disclosure risk review, and training. In the future, organizations managing restricted data may need proactively to take into consideration reducing inequalities in access to scientific data, preventing unethical use of data, and managing various types of data.

1. Introduction

Since the introduction of the Five Safes in the mid 2010s (e.g, Desai, Ritchie, and Welpton, 2016; Ritchie, 2017), many organizations managing restricted data have adopted the framework for the management of restricted and confidential data. The Five Safes framework helps organizations set guidelines for *safe data* created by data providers, *safe projects* for public good, *safe people* who are authenticated data users, *safe settings* in which data are being used, and *safe outputs* from analyzing data. The Five Safes have been well integrated throughout the data life cycle, and have led to good stewardship practices to make scientific data FAIR (Findable, Accessible, Interoperable, Reusable). It also helps multiple stakeholders balance data utilization with protection of subject privacy and data confidentiality. Despite successful implementation of the Five Safes, organizations encounter unintended challenges. In this paper, we review the current practice of restricted data management and discuss challenges and future directions, focusing on data use agreements, disclosure risk review, and training for data users.

While organizations implement multiple modes of data access (e.g., virtual data enclaves [VDE], physical data enclaves [PDE], secure encrypted file downloads), our discussion may

Key words and phrases: training, confidentiality, data governance.
Corresponding author: J. Jang, joyjang@umich.edu.

JPrivacy Confidentiality
www.journalprivacyconfidentiality.org
DOI:10.29012/jpc.844

apply mostly to VDE and PDE. Further, our discourse is centered around quantitative data, although we do not restrict the implications to only that type of data. In other words, even though our discussion on current practices may be largely reliant on our experience with quantitative data accessible via VDE or PDE, the implications of our study may extend to newly emerged data types such as research notes, video, and electroencephalography.

2. Data Use Agreements

Data use agreements (DUAs) are risk mitigation tools that clarify expectations among multiple stakeholders (O'Hara, 2020). DUAs must be entered into before any use or access to data by users, and may require periodic updates. DUAs may contain all Five Safes components: safe data (description of how data have been and will be treated for protection of any disclosure risks); safe people (data users' credentials); safe projects (research proposals demonstrating the intended data use); safe setting (plans for safe data access and handling); and safe outputs (procedures or rules on output publication and release). For some organizations, DUAs are stand-alone documents containing all five components. Other organizations require quite short DUAs accompanied by separate materials such as a detailed research proposal, approval or exemption from an Institutional Review Board (IRB), and CVs from participants in the research project. Involvement of multiple stakeholders in DUAs means that DUAs allow for negotiations and pursuit of consensus among parties.

Many organizations are bound by federal, state and local laws, regulations, or policies reflecting their capability to access direct identifiers in the datasets. DUAs specify terms and conditions for data access and use, and clarify liability issues in advance. This upfront emphasis on DUAs would help mitigate confusion regarding liability in case of data breaches or suspected security incidents. DUAs require data users' authenticated credentials; some organizations additionally ask for involvement of the researchers' institutions in DUAs as a leverage to enforce consequences for the institution (Levenstein, 2020). Not only for legal leverage, but also the involvement of institutional representatives in DUAs would help implement safe use of data by researchers. Research shows that many data users care more about their personal penalties (loss of access and funding, opinions of colleagues) rather than legal ones, if any incident happens (Green, et al., 2017). Having multiple layers of liability may safeguard data breaches or protocol violations by users. However, involvement of the institutions in the DUAs may impose a hurdle for research teams with collaborators from multiple institutions or from different countries. DUAs for research projects of this nature may have to consider heterogeneous requirements with regard to data privacy, confidentiality, and liability issues, which may cause significant delays in the process of data use.

Below, we discuss four distinctive challenges that organizations encounter with regard to restricted data management: limited opportunities of data access for certain groups of individuals; DUAs for research projects involving multiple institutions; limitations on binding laws against failure to DUA compliance; and costs to access data.

2.1. Limited Opportunities for Data Access by Certain Groups.

As described, institutional involvement may help enforce consequences for both the institution and individual researchers. Data users who are affiliated with so-called typical research institutions (e.g., universities, government agencies, research institutes) have an institutional representative involved in the DUA process, and work with organizations without substantial challenges. Most of the processes are seamless, unless stakeholders raise concerns. (Even with concerns, the most serious challenge may be a delay in the process.) However, a requirement of institutional involvement can impose an insurmountable hurdle for those without an institutional affiliation, such as freelance journalists or students without academic advisors or from institutions with no experience. Researchers and institutions negotiate details in DUAs and pursue consensus with data managing organizations, which could be a tremendous burden for small institutions. While institutional involvement is meant to help keep safe people safer, it may have unintentionally excluded researchers without that leverage. An exemption for those who have been authorized and been good users at other organizations may need to be considered, and a template agreement may that may mitigate the burdens is available (Levenstein, et al., 2018; O'Hara, 2020). Effective user training for ethical and scientific use of data may be helpful to alleviate concerns regarding data misuse by those with limited experience.

DUAs (or other supplement materials) require safe settings to access restricted data. Safe setting in DUAs designates a space in which no authorized views are allowable, for instance, an office space with a door that lacks a window. Shared space is not accepted by some organizations as a secure setting. Again, this requirement may impose a barrier for those with limited resources, such as students who would access restricted data in a shared office or cubicle. Organizations may need to consider embracing those who have limited resources by accommodating their needs (e.g., using a privacy screen for those who access data in a shared office).

2.2. **DUAs for Research Projects Involving Multiple Institutions.** When researchers from multiple institutions collaborate in a single research project, each institution would enter into the DUAs. DUAs clarify expectations and responsibilities for each institution according to the research plan. The process is often complicated when institutions are located in different countries (e.g., legitimacy of credential authentication or IRB approval in different languages). O'Hara (2020) suggests considering other forms of documentation in multi-site research projects, such as a Memorandum of Understanding (MOU) and identification of conflicts of interest. In some cases, requiring identical DUAs with all participating institutions, although requiring extensive time to complete, may reduce confusion as compare to differing DUAs across institutions. Ultimately, to streamline the process of multi-site research projects, it may be helpful for organizations to consider incentives for good data users in different projects or even in different organizations. For example, the Research Passport of the Inter-university Consortium for Political and Social Research (ICPSR) expedites access to restricted data by giving researchers credits and visibility for "safe" actions in their past experiences with restricted data (Levenstein, et al., 2018). This type of verification on users' cumulative "safe" actions would tremendously help the procedures of DUAs across multiple institutions.

2.3. **Limitations on Binding Laws Against DUA Non-Compliance.** Failure to comply with a DUA may result in immediate termination of data access and further actions that depend on the severity of the failure. Organizations establish procedures to respond to data security and breach incidents; some funders require an one-hour reporting and procedures to minimize the damage of the data breach or confidentiality disclosure. In the United States, violation of the Health Insurance Portability and Accountability Act (HIPAA) privacy standards can impose a civil monetary penalty on the individual by the Department of Health and Human Services. Organizations bound by specific laws such as HIPAA must follow the high-level legal boundary. Nonetheless, most data security incidents are unintentional or inadvertent violations of the protocol. They may pose minimal risk for subjects in datasets, and thus better be handled with effective user training. Organizations may better consider DUAs as a tool for all stakeholders to share responsibilities for data confidentiality (a community model, Green, et al., 2017), rather than the one for policing or punishing one party (a policing model, Green, et al., 2017).

2.4. Costs to Access Restricted Data.

Even marginal costs of access data can be burdensome to researchers, but are also important to consider for organizations. Data access costs include staff efforts to set up the access and to create datasets for users. The costs could unintentionally exclude some groups of researchers, such as junior scholars without research funds. Organizations and funding agencies could proactively intervene to by waiving the costs of data access for researchers with limited resources to. Doing so would help achieve Open Science (OECD, 2015)—aiming to share data with minimal barriers for all researchers from different backgrounds.

3. Disclosure Review Practices

Safe output refers to statistical products created from the restricted and/or sensitive data that are being vetted and approved as non-disclosive. Organizations help researchers utilize restricted data as effectively as possible without compromising data privacy and confidentiality. Safe output by safe people must go through a vetting process for disclosure risks. Disclosure review rules and procedures are set up in earlier steps of the data access process, such as the DUAs. Some data providers prefer to set up standards of disclosure avoidance rules and procedures with organizations in the data depositing process. Data providers and organizations also often discuss dissemination modes and tiers of access to establish the disclosure avoidance rules and procedures.

Disclosure review rules and procedures vary by types of data and access modes. Alves and Ritchie (2020) articulate two approaches to managing output-vetting: "rules-based" and "principle-based." The rules-based approach establishes a certain set of strict rules regarding disclosive information and scrutinizes research outputs created from restricted data based on the rules. On the other hand, the principle-based approach allows flexible negotiation between researchers and output vetting staff. The goal of the organizations is to implement efficient and effective procedures to protect data confidentiality and minimize disclosure risks as well as to maximize the data utilization (Griffiths, et al., 2019; Levenstein, 2019). Most organizations apply the rules-based output vetting approach, with a certain level of flexibility, to various data types.

We review below the current practice and future directions in four domains: common output vetting requirements at organizations; reviewers of statistical outputs; automatic disclosure review procedure; self-vetting that relies on "safe setting" and "safe people."

3.1. **Output Vetting Requirements.** Organizations set up a standardized procedure for output vetting, including but is not limited to output format, contents, and timeline to process each request. To illustrate, Table 1 summarizes output vetting requirements and considerations currently in place at many data archives at ICPSR. Most organizations have their own requirements and considerations in the restricted data use process. While standardizing the process and requirements could help streamline the procedures, it seems implausible due to different requirement by funders and data providers.

Table 1: Output Vetting Requirements and Considerations at ICPSR

Item	**Requirements**	**Examples**
Format	Presentation ready format required/preferred (.pdf, .docx, .xlsx).	Raw outputs from statistical packages (e.g., SAS log, Stata log-files, M-Plus log) not accepted.
Contents	A description of the sample, sub-sample, analytic approach, and definitions of variables used in the analyses. Summary statistics for variables used in the analysis. Checklist (help self-vet before sending it to the vetting staff). Supporting documents (programming files)	Minimum cell size threshold is clearly described in the output vetting instruction. Minimum cell size threshold differs by type of dataset and linkage capability.
Timeline	Depends on the output, but most vetting is completed within 10 business days.	Missing requirements, insufficient supporting documents or materials would significantly extend the timeline.

3.2. **Reviewers of the Statistical Outputs.** It is preferred that organizations have output-vetting reviewers with background in statistics or subject areas, but this is not a requirement. More important aspects are: 1) independence of the reviewers; 2) the four eyes principle; and 3) manageable workload without excessive pressure (Griffiths, et al., 2019).

Most organizations have designated individuals responsible for output vetting. For example, there are at least five experts at ICPSR all the time, with two or three back-ups, who vet outputs created from VDE or PDE. These experts are mostly ICPSR staff members who are not affiliated with any research projects of users (independence). To bolster the confidence regarding whether to release output, organizations have a group of reviewers (four eyes principle; managing workload). Some organizations operate a committee who discuss the risks of data confidentiality and privacy from research outputs. The committee usually consists of a group of experts to oversee data confidentiality and evaluate disclosure risks from the use of restricted data. For example, the ICPSR Disclosure Review Board (DRB) fills a leadership and scholarly role in the disclosure avoidance community, and serves as a decision-making body within the ICPSR with regard to disclosure risks and exceptions to existing policies. The ICPSR DRB consists of a Chair (ICPSR Privacy and Security Officer), Vice-Chair, and ten experts within and outside the organization. Individual ICPSR reviewers can query the DRB about disclosure risks on outputs and defer the approval decision to the DRB. Further, DRB reviews the ICPSR disclosure rules in light of new regulations and changes to the wider data environment, assesses new disclosure reduction methods and technologies for possible adoption, and develops rules around them. The ICPSR DRB convenes every month.

Having a group of experts (committee) who can provide a second set of eyes on disclosure risks would be beneficial with regard to confidentiality and privacy protection, but it could create frustration for data users on a tight timeline. It is important for organizations to consider the procedure of committee involvement to be flexible, e.g., an *ad hoc* subcommittee available for immediate consultation on specific requests.

3.3. **Automated Disclosure Review.** Organizations try to standardize the process of disclosure review despite disparate requirements by data type, funding agencies, and data depositors that hamper progress. High-level standardization of the disclosure review process helps streamline the vetting process, and may save the vetting timeline. In terms of vetting guidelines, standardization would be easy for the rules-based approach (setting common strict rules across datasets and organizations), but it could diminish the utilization of data if some of the output were unnecessarily determined as risky. Standardizing output vetting using the principle-based approach may be easier to implement; having a rule of thumb to vet each output and releasing if risks are negligible (Griffiths, et al., 2019). One caveat regarding standardization of the principle-based approach is that organizations may want highly-qualified expert reviewers to assess the disclosure risks of statistical outputs.

Most organizations support a pool of experts to perform disclosure risk reviews, which is often time- and resource-consuming. Instead, organizations may consider an automated disclosure review system since output checking for disclosure risks is not necessarily a statistical matter but an operational matter (Alves and Ritchie, 2020). In fact, some organizations have already implemented a machine-driven output checking for disclosure risks with regard to relatively simple matters such minimum cell thresholds, although other organizations still rely rely on human powers for the output checking. Stocchi and Bujnowska (2021) summarized the automatic Stata programming developed by Ritchie, et al. (2021), suggesting that the automated checking may work more effectively by a joint effort with expert personnel. Ritchie, et al. (2021) pointed out that automated tools may over-protect data by treating every possible case as an actual risk (which might compromise the utilization of restricted data). Also, the tool may over- or under-protect disclosure risks due to its inability to determine the context of data use (Ritchie, et al., 2021). Combination of the automated review process with expert check-ups might be most effective. Further, safe output created by safe users may help the automatic disclosure review system work the best. Organizations may invest in user training for good output preparation and checking behaviors, which eventually saves reviewers' efforts and other resources.

3.4. **Self-vetting that Relies on "Safe Setting" and "Safe People".** Outputs created within a VDE or PDE must go through a vetting process before retrieval, either by experts or by automated vetting system. On the other hand, organizations have to rely on an output self-vetting by data users who access data via a secure download method. Organizations do not scrutinize each output created from the secure download but ensure for "safe setting" and "safe people" by providing training and guidelines. Audits on data management and use in safe settings by safe people are also conducted by many organizations. However, given greater risks of disclosure with secure encrypted data download dissemination, efforts for safe data may be required.

4. Training

Organizations require user training before accessing data, which includes, but is not limited to, data confidentiality, data use procedures (steps to restricted data application, output review process), and sanctions in relation to violations of data use protocols. Training may include passive materials (print-outs or videos), interactive materials (one-on-one phone or video sessions), or quizzes. While written training materials may work better for users to follow procedures, animated and interactive materials also provide benefits in terms of translating the training into the practice (Palmiter, Elkerton, and Baguette, 1991). Combination of both passive and interactive training approaches would operate the best. Training requirements depend on types of data, funding agencies, data providers, and methods to access data (VDE/PDE or secure download); thus, user training and staff training may vary within the organization.

Recently, there has been growing consensus that user training should focus on a "community model," not a "policing model" (Green, et al. 2017). Training based on the policing model operates as a tool to make sure that researchers obey rules, assuming data users to be potential rule-breakers. On the other hand, the community model considers data users as colleagues with a shared goal of data confidentiality. (Details about the training theory are available in Green, et al. (2017).) In fact, many organizations rarely encounter substantial data breach incidents, but most of the common incidents result from inadvertent mistakes and ignorance of protocols by researchers. Effective training may better catalyze attitudinal shifts by focusing less on punishment (Green, et al. 2017).

From an organization perspective, effective training requires extensive resources. Some restricted data accessing mechanisms require yearly updates to all research materials such as DUAs, IRB approval/exemption, and training. For organizations with diverse datasets and various types of users, tracking the yearly progress for every researcher and team may be burdensome. While the community model training would work effectively, having a good facilitator may not be easy for some organizations, and updating the materials frequently may be a hurdle for many organizations. Some organizations are moving toward automated and routine training for data users and also their staff, which may resolve some issues. Also, standard training that authorizes users to access data across organizations may help reduce the burden that is imposed to organizations.

While from a user perspective it is effective to have a condensed, succinct version of training, the content of training may keep being extended. For example, there have been growing concerns for data providers and managing organizations that data are being misused. The conclusions of research where restricted data is being used are sometimes harmful to specific groups or stigmatizing to a certain group of individuals. Organizations now consider inclusion of data ethics in training materials, although how to incorporate ethics issues in a way the community model can be implemented is still in question.

5. Conclusions

In the past few decades, there have been efforts by multiple stakeholders (e.g., researchers, organizations, publishers, and funders of scientific research) to make scientific data FAIR. Technological advances such as search tools, vocabularies and infrastructures have assisted in discovery and reuse of scientific data. Many organizations have implemented the Five Safes framework in their data management to protect the confidentiality of human subjects as well as to promote reproducibility and transparency. Despite the effort, we observe that the safeguards could generate unintended challenges to certain groups of individuals (e.g., institutional approval that could exclude researchers without institutional affiliation) or in different areas (e.g., rigorous output checking that requires extensive insights from experts). This may raise questions for organizations with regard to future directions of data management with the Five Safes; for example, whether and how organizations govern the inequalities in access to scientific development and prevent unethical use of data (such as exploitation of indigenous population, group harm to underrepresented or minority groups), which is one of the essentials of Open Science (UNESCO, 2021). Furthermore, organizations now face additional challenges with newly emerged data types. Organizations may need to consider a streamlined and standardized data management while allowing for a greater degree of flexibility to incorporate such data in the future.

References

Alves, K., and Ritchie, F. (2020). Runners, repeaters, strangers and aliens: Operationalising efficient output disclosure control. *Statistical Journal of the IAOS*, 36(4):1281-1293. https://doi.org/10.3233/SJI-200661

Desai, T., Ritchie, F., and Welpton, R. (2016). Five Safes: Designing data access for research. https://www2.uwe.ac.uk/faculties/BBS/Documents/1601.pdf

Green, E., Ritchie, F., Newman, J., and Parker, T. (2017). Lessons learned in training 'safe users' of confidential data. *Work Session on Statistical Data Confidentiality.* https://pdfs.semanticscholar.org/548f/4ad0434c0f67183d557fed9661bd8baa2c07.pdf

Griffiths, E., Greci, C., Kotrotsios, Y., Parker, S., Scott, J., Welpton, R., and Woods, C. (2019). Handbook on Statistical Disclosure Control for Outputs. *Safe Data Access Professionals Working Group.* https://ukdataservice.ac.uk/app/uploads/thf_datareport_aw_web.pdf

Levenstein, M. C. (2019). Managing Research and Data for Reproducibility and Transparency. https://opremethodsmeeting.org/wp-content/uploads/2019/10/Reproducibility_Levenstein_presentation.pdf

Levenstein, M. C. (2020). Addressing Challenges of Restricted Data Access. https://hdl.handle.net/2027.42/156407

Levenstein, M. C., Tyler, A. R. B., and Davidson Bleckman, J. (2018). *The Researcher Passport: Improving Data Access and Confidentiality Protection: ICPSR's Strategy for a Community-normed System of Digital Identities of Access.* ICPSR White Paper Series, Ann Arbor, MI: University of Michigan Inter-university Consortium for Political and Social Research. https://hdl.handle.net/2027.42/143808

OECD (2015), Making Open Science a Reality. *OECD Science, Technology and Industry Policy Papers*, No. 25, OECD Publishing, Paris. https://doi.org/10.1787/5jrs2f963zs1-en.

O'Hara, A. (2020). Model data use agreements: a practical guide. In: Cole, Dhaliwal, Sautmann, and Vilhuber (eds.), *Handbook on Using Administrative Data for Research and Evidence-based Policy*. https://admindatahandbook.mit.edu/print/v1.0/handbook_ch3_DUA.pdf

Palmiter, S., Elkerton, J., and Baggett, P. (1991). Animated demonstrations vs written instructions for learning procedural tasks: a preliminary investigation. *International Journal of Man-Machine Studies*, 34(5):687-701. DOI: 10.1016/0020-7373(91)90019-4.

Ritchie, F. (2017). The 'Five Safes': a framework for planning, designing and evaluating data access solutions. Data for Policy Conference. https://doi.org/10.5281/zenodo.897821.

Ritchie, F., Green, E., and Smith, J. (2021). Automatic Checking of Research Outputs (ACRO): a tool for dynamic disclosure checks. EUROSTAT Statistical Working Paper. https://doi.org/10.2785/75954

Stocchi, M., and Bujnowska, A. (2021). Automatic checking of research outputs. https://unece.org/sites/default/files/2021-12/SDC2021_Day2_Stocchi_AD.pdf

UNESCO (2021). *UNESCO Recommendation on Open Science.* Paris: UNESCO, 2021. https://doi.org/10.54677/MNMH8546