disc
Data Integration Support Center
WestEd

2023
Issue 4

Protecting the Privacy of Personal Information in Integrated Data Systems: Disclosure Avoidance Methods

INTEGRATED DATA SYSTEMS

connect data over time and across sectors to provide data insights that support leaders in answering policy questions, directing resources, and better supporting individuals.

Integrated Data Systems (IDSs) link individual-level data from multiple agencies, such as schools, juvenile justice agencies, and human services agencies, often with a focus on children. The systems can be used for a variety of purposes, including case management or program monitoring and evaluation. By linking data across public agencies, redundant data collection can be minimized, while the ability to utilize the information to improve programs and services for the community is maximized.

While IDSs have numerous benefits, there is also an increased possibility of disclosure when data sets are shared and/or integrated across different sectors and/or public agencies (Actionable Intelligence for Social Policy, n.d.; Nelson & Zanti, 2020). Disclosure occurs when confidential information about an individual, organization, or entity is revealed in the data set. In this paper, we refer to *disclosure* as encompassing any of the following:

- **identity disclosure**: when a subject's identity can be identified in a data record containing confidential information
- **attribute disclosure**: when a sensitive attribute of a subject's identity is revealed from a data release
- **inferential disclosure**: when released data make it more likely that the consumer of the data release can identify or determine an attribute of a subject in the data

The process used to protect the confidentiality and privacy of personal information within an integrated data system is called **disclosure avoidance**. This

document provides information on disclosure avoidance within the context of an IDS, including the following:

- the risks of identity disclosure in an IDS
- best practices to reduce the risk of identity disclosure
- considerations on disclosure-avoidance strategies during the life cycle of an IDS
- additional resources on disclosure avoidance and IDSs

Disclosure Avoidance in the IDS Life Cycle

An IDS provides valuable population-level information that can make it easier for organizations to work together to improve social services for their communities. Within the power of an IDS comes inherent privacy risks that should be addressed within the IDS's life cycle. The primary goal for parties who are linking, analyzing, and reporting on data sets should be to minimize privacy risks at each stage of the IDS development and reporting process. They should seek to strike a balance between protecting confidential information about the data's subjects while also providing the most utility possible for the IDS (Griffiths et al., 2019).

Agencies and interested parties seeking to develop an IDS have an obvious legal and ethical obligation to protect each individual's privacy and to minimize disclosure. Federal agencies and states are increasingly including stringent privacy and confidentiality standards to ensure that the possibility of disclosure is limited. This also means that organizations who seek to link their data must collaborate to ensure that their strategies for integration meets their respective privacy and confidentiality standards (National Academies of Sciences, Engineering, and Medicine, 2017a). Disclosure avoidance methods are equally important because everyone deserves to know that sensitive data collected about them will be kept confidential. The success of an IDS depends on its "social license," which includes credibility and public trust (Finch et al., 2018). If people do not trust public agencies with their data, data quality and response rates will decline, making it more difficult to reap the benefits of integrated data (Hundepool et al., 2010; Schmutte & Vilhuber, 2020).

To be successful, disclosure avoidance practices need to be identified and implemented throughout each stage of the IDS. Although the process of implementing an IDS is complex, this paper focuses on the following four distinct stages of IDS implementation:

- data linkage
- de-identification
- data access
- data reporting

The sections that follow discuss the best practices of disclosure avoidance in each stage.

Data Linkages Across Sectors

Public agencies seeking to link their data must make decisions about how data will be linked and matched across data sets. A strong data governance structure supports multiagency efforts to identify the availability and sources of the data, make decisions about the data in the IDS, and minimize identity disclosure during the data-linking process. Many public agencies rely on direct identifiers for subjects in their systems to mask identifying information. However, some unique identifiers (e.g., social security numbers) may still be considered personally identifiable information and need additional protections to avoid potential disclosure, particularly as data are linked across sources in an IDS. Agencies should begin the

data-linkage process by considering the following questions about the prospective data to be linked from their respective agencies:

- Why do these data need to be protected? (Hundepool et al., 2010)
- Are there legal or institutional guidelines that must be followed? (Schmutte & Vilhuber, 2020; Templ et al., 2021)
- What are the ethical considerations? (Schmutte & Vilhuber, 2020)
- What privacy concerns does the community have? (Bowen et al., 2021; Finch et al., 2018)

The best practice to reduce disclosure is a full-separation approach in which the data linkage processes are fully separated from the data analysis processes. In this approach, only those staff who conduct the data linkage processes have access to the direct identifiers (e.g., full name, date of birth) without the rest of the data. Those conducting the data analysis only have access to the de-identified, coded data (Harron et al., 2017). The direct identifiers are then stored separately from the rest of the data and only utilized to link new data or refresh existing data.

To reduce the possibility of disclosure, agencies should consider the following questions before data are linked:

- Are direct identifiers collected with the data sets that will be linked across the public agencies? (Garfinkel, 2016)
- How will direct identifiers be separated from data used for analyses?
- What analyses will the de-identified linked data support?

De-identification of Linked Data

In general, to avoid sharing data that contain private information on individuals, public agencies only release data that have been de-identified. It is important to note that de-identification is typically accomplished and evaluated through software that uses complex mathematical techniques. Most de-identification practices remove direct identifiers (e.g., names, phone numbers, social security numbers) and replace them with quasi-identifiers, which mask the direct identifiers and limit the possibility of disclosure (Polonetsky et al., 2016).

However, when an IDS links data across multiple sectors, there is an increased risk of de-identified data becoming "reidentified." That is, multisector data linkages can result in additional personal attributes being revealed that can be used to identify a subject in the data set (Garfinkel, 2016).

Noise infusion. To avoid possible disclosure, an alternative to the de-identification method is a method known as *noise infusion*. Noise infusion involves using the linked data sets to create a new (sometimes called "synthetic") data set that adds "noise" to specific cells that have a high probability of disclosure (Garfinkel, 2016). Noise infusion can include techniques such as blanking and imputing and record swapping. Blanking and imputing involves removing highly identifying values in a data set and replacing them with random values. Record swapping occurs when attributes are swapped between subjects in the data set (National Academies of Sciences, Engineering, and Medicine, 2017b). Agencies may also consider combining quasi-identifiers with noise infusion to further reduce the risks of identification.

Once public agencies de-identify the linked data for further analyses, they can consider the following questions:

- Are there other data elements that can be used to identify a subject?
- What are the risks posed to subjects if identification occurs? (Garfinkel, 2016)

Data Access

Public agencies and other data contributors will need to define and implement procedures for data access. Frequently, only de-identified data are made available to qualified researchers under legally binding data-use agreements (Polonetsky et al., 2016). These agreements limit what researchers can do with the data. For instance, these agreements generally prevent researchers from being able to attempt identification of subjects in the data, further link data to other data sets, or redistribute the data. Other options for managing data access involve query capabilities that are more sophisticated and can be useful for IDSs that intend to limit raw, de-identified data releases. These options include allowing researchers to submit queries through which the data system provides the results rather than allowing researchers direct access to the linked data.

When considering how to manage data access for an IDS, the public agencies and others involved should consider the following questions:

- Who will have access to the data in the IDS? How will access be granted, and who will determine the appropriate access? What agreements are needed to access the data? (Public agencies should consider how their governance supports appropriate access and use of the data.)
- Will access be granted to the identifiable data and for what purposes? Will access only be granted to the de-identified data? What query capabilities will the IDS support, if any?
- Are there mechanisms in place to inform the agencies if there is an attempt to reidentify the de-identified data set? (Garfinkel, 2016)

Data Reporting

Once data are released and are actively being used by researchers, public agencies should continue to implement disclosure avoidance methods. An IDS can impose rules on the reporting of the data by the IDS and/or by individuals who can access and report the data as IDS data users. If the IDS releases data reports to the public, methods such as aggregate data tables can remove much of the disclosure risk. However, some risk remains when one or more subjects have unique characteristics. This scenario most often shows up in small sample sizes in which data suppression of both sensitive and nonsensitive data may be

By thoughtfully considering disclosure avoidance throughout the development of an IDS, the agencies that are linking their data can start to develop the most appropriate disclosure avoidance strategy.

necessary to ensure disclosure does not occur. Data governance can also require specific rules around suppression for data users. In most government agencies, a cell size of 3 or smaller warrants suppression, though some may require any cell size of 10 or smaller to be suppressed during reporting.

IDSs should consider the following questions in relation to data reporting:

- Are there minimum cell size requirements for the public agencies that provide data to the IDS?
- What is the process for reviewing the results of data reports prior to releasing them to the public?
- What is the minimum risk level of disclosure that the agencies involved in the IDS are willing to accept to reap the benefits of the IDS? (Schmutte & Vilhuber, 2020; Templ et al., 2021)

Developing an IDS Disclosure Avoidance Strategy

There are many methods of disclosure avoidance but few hard-and-fast rules or guidelines. For instance, there is no shared definition of what it means to be "private enough," and privacy researchers have increasingly moved away from a binary evaluation of disclosure (e.g., "Was an individual identified or not?") and toward a cumulative risk of disclosure (Matthews & Harel, 2011).

By thoughtfully considering disclosure avoidance throughout the development of an IDS, the agencies that are linking their data can start to develop the most appropriate disclosure avoidance strategy. Regardless of the methods chosen, it is important to be transparent around the disclosure decisions and risk assessment process, as transparency increases trust and allows for replication of data analysis and error checking (Hundepool et al., 2010). One way to increase transparency is to engage the community that is correlated to the data. It is likely that the subjects in the data sets have unique privacy concerns and opinions about how their data are represented, analyzed, and released (Brown et al., 2021). As Actionable Intelligence for Social Policy notes, IDSs face several unique hurdles to public trust. Meaningful engagement and communication with interested parties as well as strong privacy protections will increase the credibility of the IDS (Finch et al., 2018). Additionally, Finch et al. (2018) recommend the inclusion of a diverse set of interested parties in the discussions and decisions about disclosure avoidance, as privacy risks vary across the community.

How the Data Integration Support Center Can Help

The Data Integration Support Center (DISC) at WestEd can support public agencies' ongoing privacy efforts. Some of the forms of technical assistance that can be provided by DISC include the following:

- assisting in the design of an IDS's architecture to support privacy at each stage of the IDS life cycle
- performing a statistical analysis of multiagency linked data and recommending best approaches for performing statistical avoidance measures, with utility versus privacy in mind
- reviewing statistical-avoidance methodologies and providing recommendations for improvements
- curating documents, resources, and examples of thoughtful disclosure avoidance methodologies
- developing a playbook to assist staff in sufficiently implementing statistical avoidance methodologies

REFERENCES

Actionable Intelligence for Social Policy. (n.d.). *Centering racial equity throughout data integration*.

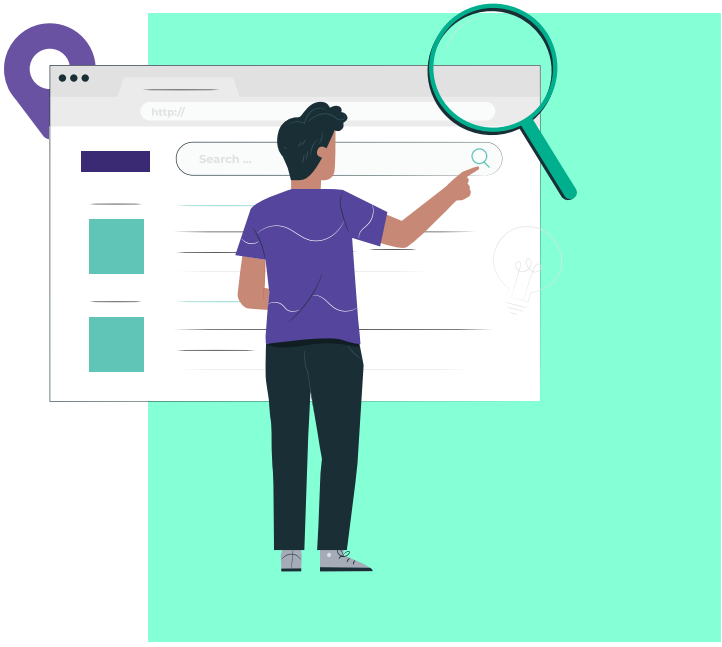
Bowen, C., Narayanan, A., & Scally, C.P. (2021). *Using differential privacy to advance rural economic development: Applying data privacy and confidentiality methods to industry employment data*. Urban Institute. https://www.urban.org/research/publication/using-differential-privacy-advance-rural-economic-development

Brown, K. S., Su, Y., Jagganath, J., Rayfield, J., & Randall, M. (2021). *Ethics and empathy in using imputation to disaggregate data for racial equity: Landscape scan findings*. Urban Institute. https://www.urban.org/sites/default/files/publication/104678/ethics-and-empathy-in-using-imputation-to-disaggregate-data-for-racial-equity_0.pdf

Finch, K., Hawn Nelson, A., Jenkins, D., Burnett, T. C., Oliver, A., & Martin, R. (2018). *Nothing to hide: Tools for talking (and listening) about data privacy for integrated data systems*. Future of Privacy Forum; Actional Intelligence for Social Policy. https://www.aisp.upenn.edu/wp-content/uploads/2018/10/FPF-AISP_Nothing-to-Hide-FINAL.pdf

Garfinkel, S. (2016). *De-identifying government datasets* (No. 800–188; NIST Special Publication). National Institute of Standards and Technology.

Griffiths, E., Greci, C., Kotrotsios, Y., Parker, S., Scott, J., Welpton, R., Wolters, A., & Woods, C. (2019). *Handbook on statistical disclosure for outputs*. The Heath Foundation.



Harron, K., Dibben, C., Boyd, J., Hjern, A., Azimaee, M., Barreto, M. L., & Goldstein, H. (2017). Challenges in administrative data linkage for research. *Big Data & Society*, *4*(2), 1–12. https://doi.org/10.1177/2053951717745678

Hawn Nelson, A. L., & Zanti, S. (2020). A framework for centering racial equity throughout the administrative data life cycle [Special issue]. *International Journal of Population Data Science, 5*(3), 1367. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8110889/

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz, R., Naylor, J., Nordholt, E. S., Seri, G., & De Wolf, P. P. (2010). *Handbook on statistical disclosure control: Vol. 1.2*. A Network of Excellence in the European Statistical System in the Field of Statistical Disclosure Control.

Matthews, G. J., & Harel, O. (2011). Data confidentiality: A review of methods for statistical disclosure limitation and methods for assessing privacy. *Statistics Surveys*, *5*, 1–29. https://doi.org/10.1214/11-SS074

National Academies of Sciences, Engineering, and Medicine. (2017a). *Federal statistics, multiple data sources, and privacy protection: Next steps*. National Academies Press. https://doi.org/10.17226/24893

National Academies of Sciences, Engineering, and Medicine. (2017b). *Innovations in federal statistics: Combining data sources while protecting privacy*. National Academies Press. https://doi.org/10.17226/24652

Polonetsky, J., Tene, O., & Finch, K. (2016). Shades of gray: Seeing the full spectrum of practical data de-identification. *Santa Clara Law Review*, *56*(3), 593.

Schmutte, I. M., & Vilhuber, L. (2020). Balancing privacy and data usability: An overview of disclosure avoidance methods. In S. Cole, I. Dhaliwal, A. Sautmann, & L. Vilhuber (Eds.), *Handbook on using administrative data for research and evidence-based policy* (pp. 145–172). Abdul Latif Jameel Poverty Action Lab; Massachusetts Institute of Technology. https://admindatahandbook.mit.edu/book/v1.1/index.html

Templ, M., Meindl, B., & Kowarik, A. (2021). *Introduction to statistical disclosure control* (SDC). Data Analysis OG.

Suggested citation: Terrell, J., & Tiderman, L. (2023). *Protecting the privacy of personal information in integrated data systems: Disclosure avoidance methods*. WestEd.