NATIONAL ACADEMIES Sciences Engineering Medicine

Consensus Study Report
Issue Brief

A Roadmap for Disclosure Avoidance in the Survey of Income and Program Participation

What Survey and Database Managers Can Learn

In 2022–2023, the Committee on National Statistics within the National Academies of Sciences, Engineering, and Medicine convened a panel of experts to examine disclosure avoidance in the Survey of Income and Program Participation (SIPP). Following are some of the key implications of the report for database managers.

SHOULD I BE DOING ANYTHING DIFFERENTLY REGARDING DATABASE PRIVACY?

The times are changing. Once data were highly dispersed, sometimes requiring access to physical records in multiple locations. With data increasingly becoming digitized and becoming available online, data intruders have ready access to a wide variety of data, and the amount and types of such data continue to grow. Further, advances in computing technology, software tools, and internet search capabilities all have combined to enable the processing of large amounts of data from diverse sources in ways that might allow the identification of survey respondents. At the same time, new approaches for providing data access while also protecting data are also being developed, and these can be incorporated into new strategies for protecting confidentiality.

The panel found that the disclosure risks faced by SIPP are likely to be high and need to be assessed rigorously building on the initial work conducted by the Census Bureau. In the case of SIPP, the initial work may underestimate risk since not all of the variables and databases open to data intruders were considered, risks were estimated only for primary respondents, and analysts did not consider the full dimensionality of the data, in which both data on other members of the household and data on change over time might be used to help identify respondents. For example, interracial marriages are relatively uncommon, as also are large age differences between spouses; thus, knowing that a black male is married to an Asian female who is 12 years younger helps to quickly narrow the number of possible matches.

HOW SHOULD WE BE EVALUATING DISCLOSURE RISK?

The panel concluded that measuring both relative risk and absolute risk is important (Conclusion 3-1). Relative risk may be low while absolute risk is high; conversely, if absolute risk is low, then relative risk might be allowed to be high.

It is important to consider all of the data that might be used by a data intruder, which may require examining multiple databases (Conclusion 3-4). For household surveys, also consider how information about other members of the household may affect disclosure risk. For longitudinal surveys, also consider how information about change over time may affect disclosure risk.

ALLOWANCES FOR USER NEEDS

Different surveys have different users, and user needs may vary from one survey to another. In the particular

case of SIPP, which is a highly complex database, data users tend to be highly sophisticated, conducting complex analyses and making use of multiple modules within SIPP. These needs set constraints on what types of disclosure avoidance approaches might be workable. For example, reducing the public use file to a set of core variables would not be workable, and unless a table generator or remote analysis platform is highly sophisticated, it will not meet most needs of current users (though it might increase the number of SIPP data users). For other surveys, a table generator might meet many user needs, and might also be simpler to develop (given the complex file structure used in SIPP).

The panel concluded that providing a highly comprehensive data file is needed for SIPP users, making the availability of a public use file very important (Recommendation 9-3), along with readily accessible means of conducting those types of analysis that a privacy-protected public-use file cannot support (Recommendation 9-4).

METHODS FOR MAKING DATA AVAILABLE

A key strategy, consistent with recent legislation, is to make the data available through multiple tiers of access (Recommendation 4-3). Currently, SIPP is only available through two modes of access: a public-use file and access through a Federal Statistical Research Data Center (FSRDC). Future re-identification studies may find that additional limits will be needed on the public-use file, but relying on FRSDCs to be the sole alternative to a public-use file would be burdensome on researchers and create inequities of access (Conclusion 2-3). Thus, there is a need for an intermediate tier of access that will offer fuller access to the data without imposing all of the restrictions that are associated with FSRDCs. Secure online data access seems especially well suited for this need (Recommendation 5-1).

Federal agencies might also consider whether their current data dissemination approaches are consistent with recent legislation (i.e., the Evidence Act and the Information Quality Act) that promotes data access and provide multiple tiers of access. Procedures that were established prior to these acts might be modernized to better accommodate them (Recommendation 9-5).

INVOLVEMENT OF OUTSIDE RESEARCHERS

Researchers outside the Census Bureau can bring new perspectives and expertise to aid in identifying and handling disclosure risks (Recommendation 3-3).

THE CURRENT STATE OF TECHNOLOGY

Some powerful tools for disclosure avoidance do not yet have the capacity for handling databases with the size and complexity of SIPP. These limitations both constrain what options are available for protecting SIPP respondents and create a research agenda for further work.

- Synthetic data have been used in SIPP to allow the merging of administrative data with SIPP data while protecting privacy, initially with files in which selected data were synthetic, and ultimately with files in which all data are synthetic. However, these files contained only a limited number of SIPP variables. Current technology would support replacing a limited number of variables in the public-use file with synthetic data, but not the creation of a fully synthetic and complete dataset (Conclusion 6-9; Recommendation 6-4).

- Differential privacy is another tool that might be useful in limited applications, such as within a table generator, but current technology is not well enough developed to support creating a complete SIPP data file with differential privacy.

- The technology for providing secure online data access exists and has been used successfully by multiple agencies and for multiple datasets (Recommendation 5-1). However, in order to limit the burden on Census or its contractors in supporting data validation and to disclosure review, there would be value in creating automated systems to support such work.

THE IMPORTANCE OF TRANSPARENCY

The use of disclosure avoidance techniques affects which modes of access will be most appropriate for researchers, what analytic techniques may be applied, and how the data are interpreted. It is important for Census Bureau communications to support such needs (Recommendation 3-2).

FOR MORE INFORMATION

This Consensus Study Report Issue Brief was prepared by the Committee on National Statistics and based on the Consensus Study Report *A Roadmap for Disclosure Avoidance in the Survey of Income and Program Participation* (2023). The study was sponsored by the U.S. Census Bureau. Any opinions, findings, conclusions, or recommendations expressed in this publication do not necessarily reflect the views of any organization or agency that provided support for the project. The Consensus Study Report is available from the National Academies Press, (800) 624-6242 or https://www.nap.edu/catalog/27169.

Division of Behavioral and Social Sciences and Education

NATIONAL ACADEMIES *Sciences Engineering Medicine*