Allocating Microdata from the National Longitudinal Survey of Youth Among Access Tiers: A Framework for Decision-Making and Initial Investigations

By Alison Aughinbaugh, Keenan Dworak-Fisher, Donna Rothstein, and Julie Yates

Introduction

The National Longitudinal Surveys (NLS), sponsored by the U.S. Bureau of Labor Statistics, are nationally representative surveys that follow the same sample of individuals from specific birth cohorts over time. The currently active surveys, which began in 1979 and 1997, collect data on samples of approximately 10,000 and 9,000, respectively, on a wide range of topics, including labor market activity, schooling, fertility, program participation, health, and much more. These data are made available to researchers in microdata form, using a tiered system of data access that includes an extensive public-use dataset and two restricted-use datasets that are only made available to researchers who go through a thorough application process and sign a written agreements making them official agents of the Bureau of Labor Statistics. Systems of tiered access like this are used with increasing frequency to accommodate the needs of data users for valuable research projects while minimizing the risks of privacy harms to data subjects. Harnessing the principle of use limitation, enhanced use of tiered access systems holds promise for simultaneously improving data access empowering research and improving privacy.

NLS is currently preparing to develop a new cohort, which is projected to begin collection in 2026 for a sample of approximately 17,000 youths aged 11 to 16. Developing this new cohort offers the NLS program the opportunity to reconsider how it structures its tiered data provision, including what conditions to place on access to data on different tiers and how to allocate different data elements among tiers. In this study, we document an initial investigation into how such a reconsideration should be carried out. We first describe the current practice of the NLS program and its foundations. We then provide a framework for how the program could reinforce these foundations and describe the criteria that the program should consider while allocating data elements among tiers. Finally, we walk through some examples of how the framework could be applied to a few categories of NLS microdata and discuss potential outcomes for the program.

Current NLS Practice and its Foundations

The NLS collects a wide variety of what is considered Personality Identifiable Information (PII): that is, any representation of information about an individual maintained by the NLS that permits the identity of the individual to whom the information applies to be reasonably inferred by either direct or indirect means. NLS-specific examples include, but are not limited to, education, financial transactions, and medical, criminal, or employment history, and information which can be used to distinguish or trace an individual's identity, such as their name, social security number, date and place of birth or death, mother's maiden name, etc., including any other personal information which is linked or linkable to an individual.

NLS is committed to maintaining the privacy of its respondents' personally identifying information. NLS collects data only for statistical purposes and not for any administrative of enforcement uses. With

respect to the collection, use, and disclosure of personal information, NLS makes every effort to comply with applicable federal law, including the Privacy Act of 1974, Confidential Information Protection and Statistical Efficiency Act (CIPSEA), the Paperwork Reduction Act of 1995, and the Freedom of Information Act (FOIA), as well as other legislation and OMB guidance. To best describe current practices of the NLS program and other BLS programs, we begin by reviewing the Privacy Act of 1974 and the Confidential Information Protection and Statistical Efficiency Act (CIPSEA).

The Privacy Act of 1974

The purpose of the Privacy Act is to balance the government's need to collect and disseminate data and maintain information about individuals, in our case NLS respondents, with the rights of respondents to be protected against disclosure of personally identifying information (PII) such as names, addresses, and social security numbers. The Privacy Act protects the rights of respondents by ensuring the confidentiality of their personal data, giving respondents opportunities to correct their data, and limiting data misuse. This confidentiality protection includes the fact of a respondent's participation in a survey.

One major effect of the Privacy Act on respondent confidentiality is the requirement for informed consent. Before an individual's personal information is collected by a federal agency, or as in the case of NLS, its contractors, respondents must be informed of the purpose of the collection, the authority under which it is being conducted, and the potential uses of their data. Informed consent helps to ensure that respondents are aware of their rights and the measures in place to protect their privacy, leading to greater trust in the research process and a willingness to participate.

The Privacy Act directly relates to the way personal information is handled during data collection as well as analysis. It requires Federal agencies to "establish appropriate administrative, technical, and physical safeguards to ensure the security and confidentiality of records and to protect against any anticipated threats or hazards to their security or integrity which could result in substantial harm, embarrassment, inconvenience, or unfairness to any individual on whom information is maintained." Such safeguards include rules and procedures governing access to and disclosure of personal data. By adhering to these principles, researchers can ensure that the personal information of respondents remains secure and confidential, thus promoting a positive relationship between researchers and participants.

At the start of each data collection, NLS respondents must agree to an NLS Privacy Act consent statement before the interview can begin. The consent statement affirms that PII will be held in confidence and not released to the public without consent and that personnel associated with the survey have signed a legal document where they pledge to protect the confidentiality of the respondents.

The Confidential Information Protection and Statistical Efficiency Act (CIPSEA)

CIPSEA is a U.S. federal law first enacted in 2002 that plays a critical role in protecting the confidentiality of respondents who provide information to federal statistical agencies. Under CIPSEA, statistical agencies are required to assure respondents that their information will be kept confidential and used only for statistical purposes. This assurance helps build trust between the respondents and the agencies,

encouraging participation in surveys and other data collection efforts. CIPSEA mandates that federal statistical agencies implement strong data security measures to protect the confidentiality of the information they collect. This includes restricting access to confidential data only to authorized personnel who have sworn to protect the information and are subject to fines or imprisonment for any violation. CIPSEA imposes strict penalties on individuals who knowingly disclose confidential information, with fines of up to $250,000, imprisonment for up to five years, or both.

Overall, CIPSEA strengthens respondent confidentiality by providing legal protections, requiring confidentiality assurances, and mandating data security measures. These provisions work together to help maintain trust between respondents and federal statistical agencies and ensure the continued availability of high-quality data for policymaking, research, and decision-making.

OMB issued extensive guidelines in 2006 to govern CIPSEA implementation, which BLS closely follows. A key element of the guidelines is their articulation of the "CIPSEA Pledge," which is provided to respondents by statistical agencies and units for collections that are bound by CIPSEA and may *not* be made under any other circumstances. The pledge, which signifies the trustworthiness of the agency's commitment to privacy, is provided in writing in the mailed advance letter or read to the respondent at the beginning of interview if the respondent did not receive the mailing. This pledge states PII is fully protected by law, data will be used for statistical purposes only, that only authorized persons working as BLS employees or as sworn BLS agents can access PII for approved official purposes, and there is a monetary fine for data misuse. The NLS version of the CIPSEA pledge also briefly describes the different files made available to researchers; for example, this pledge makes clear NLS has tiered access with a publicly available file (PUF) in addition to restricted use files with increased geographic information.

Data Access: Public-use and Restricted-use Data Files

1. *Public Use File (PUF)*

The NLS program uses a tiered system of data access for public-use and restricted-use data files. Public-use data are available for free with no registration requirement via the NLS Investigator webpage found at https://www.nlsinfo.org/investigator/. The NLS Investigator data download application allows researchers to search for variables and extract microdata. This PUF includes only de-identified data that cannot be used alone or in combination to disclose respondent identity; all PII variables are removed and not even NLS National staff have access to respondent identifying information like names, addresses, or phone numbers. Only contractors working in fielding and data archivists have PII access.

Consistent with the guidance provided by the Federal Committee on Statistical Methodology (FCSM), NLS limits the geographic detail available on the PUF, limits the number and detailed breakdown of its categorical variables, and employs a variety of statistical disclosure limitation (SDL) methods to further ensure variables cannot be used to identify respondents. SDL methods currently in use include truncation of extreme codes for certain variables (top and bottom-coding), recoding of continuous variables into intervals, rounding, banded answers, and outlier suppression.

Periodically, the NLS program undergoes a variable audit so staff become aware of new ways variables can be combined, potentially leading to further variable suppression. In 2020, a risk assessment performed by Westat searched for potential combinations of categorical variables in the NLS public-use

data that could possibly be unique in the population. Westat also analyzed NLS continuous variables and searched for additional risks that could present through matches to available external data. It concluded that the re-identification risk in the NLS public use data was low.

2. Restricted-Use Files

The NLS program has two restricted-use data files, both of which require project application and approval. The first is the NLS "geocode" file, which contains variables such as state and county of residence, college UNITID code, and more **(see Table 1)**. Geocode data are available to researchers via a Virtual Data Enclave (VDE). The second is the zip code/census tract file **(see Table 2)**, which is only available for on-site use at the BLS or at a Federal Statistical Data Research Center (FSDRC).

Most NLS variables are available on the public-use file. However, sensitive variables, such as those related to location of residence, are placed into one of the two restricted-use data files. When possible, the NLS program places a highly categorized version of the restricted-use variable onto the public-use file. **Table 3** displays a few examples of variables that are detailed in the geocode data file, but highly categorized in the public-use data file. For example, state of respondent's residence at the interview date is available on the geocode file and the variable is collapsed into four census regions in the public-use data. As another example, the geocode file contains college UNITID, which can be linked to data sets that contain characteristics of the college. The public-use file contains a college ID that allows a researcher to tell that the college is the same for a particular person across rounds, but the ID is not useful for linking to outside data and is also not comparable across respondents.

NLSY97, NLSY79, and NLSY79 Young Adult geocode data are considered controlled unclassified information and are available within a VDE. To protect the confidentially of respondents, BLS only grants access to geocode files for researchers in the U.S. who agree in writing to adhere to the BLS confidentiality policy and whose projects further the mission of BLS and the NLS program. Applicants must provide a clear statement of their research methodology and objectives and explain how the geocode data are necessary to meet those objectives. Researchers and their institutions must enter into a legal agreement called a Letter of Agreement (LOA) which requires researchers to use data only for statistical purposes with monetary fines if they violate the agreement. Part I. of the LOA states: *The data will be used only in aggregated multivariate statistical analyses for a research project specified in Section IV of this agreement. The BLS will not provide any personal identifiers*." Part VII lists 15 responsibilities of BLS agents, including complying with all laws that affect PII, while Section VIII lists 8 security provisions, again stressing *"reviewing all laws applicable to confidentiality and data provided under this agreement."* Those who are granted access to NLSY geocode data may access the VDE from approved locations on the physical premises of their institution.

Applicants for NLSY geocode data must first complete the Standard Application Process (SAP) through the online portal at ResearchDataGov (RDG). RDG, established to fulfill Section 3583 of the Evidence Act, is a portal for requesting access to restricted microdata from Federal statistical agencies. If the standard application is approved, applicants complete the BLS VDE Confidential Data Access Security Information Form where they enter information about who will sign the Letter of Agreement (LOA), and the proposed locations at their institution where researchers will access data. The NLS then sends a Letter of Agreement to be signed by an official authorized to sign the agreement on behalf of the university or institution. Enclosed with the Letter of Agreement are copies of the BLS agent agreement to be signed by each researcher, advisor, and anyone else named in the application who will have access

to the NLSY Geocode data. NLSY Geocode Letters of Agreement typically last one year for students and two years for faculty members but may be extended upon request. Researchers are required to complete annual confidentiality training or their access to data is removed. Once approved, researchers are given a VDE Non-disclosure Review Guide and accompanying Non-disclosure Review Checklist, which provide standards for researcher VDE output. The researcher must make sure that the output requested adheres to the terms and submit the completed checklist with the output. Items on the checklist include that the output contains no microdata, no small sample sizes, no revealing of geographic areas under a certain size, formatted and clearly labeled tables and charts, and more. The NLS program reviews the output for disclosure risk and must approve the output before it is released to the researcher.

For researchers who want to measure smaller geographic areas, the NLSY79 and NLSY97 surveys have restricted-use zip code and census tract files. These confidential files are available for use only at the BLS National Office in Washington, DC, and at FSRDCs on statistical research projects approved by BLS. To access a FSRDC researchers must also be able to obtain Special Sworn Status (SSS) from the Census Bureau.

Only individuals affiliated with certain organizations (U.S. Federal Agency, U.S. institution of higher education, etc.) are eligible to apply for access to these data https://www.bls.gov/rda/data/eligibility-and-access-modes.htm Like with the geocode access, applicants must complete the Standard Application Process (SAP) through the online portal at ResearchDataGov (RDG) and complete and sign the Letter of Agreement. The applicant provides a proposal that contains a detailed description of the project, and demonstrates the need for these data, that the project is of scientific merit, and that it helps to fulfill the mission of the BLS and NLS. The proposal goes through multiple layers of review within BLS, which can take several months. Once approved, a Letter of Agreement is created between the applicant's institution and the BLS; a BLS agent agreement is also created to be signed by each individual named in the application who will have access to the Zip Code and Census Tract data. NLSY Zip Code and Census Tract File Letters of Agreement typically last one year for students and two years for faculty members but may be extended upon request.

Once approved researchers receive a Non-disclosure Review Guide and accompanying Non-disclosure Review Checklist, which provide standards for researcher output. The BLS reviews the researcher output for disclosure risk and must approve the output before it is released to the researcher.

Reconsidering the Framework for Building the NLS Tiered Data System

These past and current practices have been developed over time, grounded in legal compliance, practical wisdom, and the mission of the BLS. They have incorporated new laws and regulations as well as new knowledge about the technological environment as needed, but they don't specifically describe an approach to guide the program's classification of microdata among access tiers. As the NLS program prepares for development of a new cohort of youth and new or updated systems to support it, such a framework would be helpful. In this section, we lay out a framework that the NLS program could use to guide its analyses. The framework is structured around two, broad goals: 1) complying with the relevant legal requirements; and 2) promoting the confidence of survey respondents and other data providers.

Legal Requirements

As described above, BLS adheres closely to the legal requirements imposed by the Privacy Act to "protect against any anticipated threats or hazards to their security or integrity which could result in substantial harm, embarrassment, inconvenience, or unfairness to any individual on whom information is maintained" and by CIPSEA to "prevent the identity of the respondent… [from being] reasonably inferred by either direct or indirect means." In addition, BLS collections adhere to the Paperwork Reduction Act, which reinforces CIPSEA and the Privacy Act but also requires agencies to "minimize the paperwork burden . . . resulting from the collection of information by or for the Federal Government" and "ensure the greatest possible public benefit from and maximize the utility of information created, collected, maintained, used, shared and disseminated by or for the Federal Government." Jointly applying the mandates of these laws requires balancing the imperatives of preventing privacy harms and empowering valuable uses of acquired data.

These joint mandates are exemplified in an addition to CIPSEA introduced in the Foundations of Evidence-Based Policymaking Act of 2018 (the "Evidence Act").[1] Section 3563 of the Evidence Act codifies the Fundamental Responsibilities of Statistical Agencies to: a) produce and disseminate relevant and timely statistical information; b) conduct credible and accurate statistical activities; c) conduct objective statistical activities; and d) protect the trust of information providers by ensuring the confidentiality and exclusive statistical use of their responses. Although the regulations to implement these requirements have not yet been promulgated, we can look to the prior guidance from OMB[2] to surmise their likely direction. ADD DISCUSSION

Two other additions to CIPSEA inherent in the Evidence Act are important to note. Section 3582 requires statistical agencies to expand access to their data assets while protecting those assets from inappropriate access and use. According to the law, the forthcoming regulations will require agencies to determine the "sensitivity level" of each data asset and assign assets to accessibility tiers accordingly. Such determinations, which are to be made public, may be affected by the extent to which data may be obscured, e.g., through the application of SDL methods. Although these regulations have not yet been promulgated, we can look to the reports of the Federal Commission on Evidence-Based Policymaking (CEP) and the Advisory Committee on Data for Evidence Building (ACDEB), as well as existing OMB guidance, to understand the thinking that may underlie them. The CEP recommended that data access decisions should be calibrated according to the expected public benefits of the access, the sensitivity the data, and the risk to confidentiality, adopting a concept of privacy risk to which we will return below. Similarly, the ACDEB has recommended that access tiers be allocated based on joint assessments of the data's utility and risk. OMB guidance further enunciates that "all users (must) have equitable and timely access to data that are disseminated to the public." [3]

---

[1] Note: Title III of the Evidence Act, itself entitled "The Confidential Information Protection and Statistical Efficiency Act of 2018," reiterated and enhanced the 2002 version of CIPSEA and allowed for its codification.

[2] "Statistical Policy Directive No. 1: Fundamental Responsibilities of Federal Statistical Agencies and Recognized Statistical Units" (*79 FR 71609*) describes the fundamental responsibilities and the underlying principles upon which they are based.

[3] "Statistical Policy Directive No. 4: "Release and Dissemination of Statistical Products Produced by Federal Statistical Agencies" (*72 FR 42266*).

Section 3581 of the Evidence Act encourages the incorporation of Federal administrative data into statistical datasets by creating a presumption that such administrative data are available to be shared with statistical agencies. As with the other sections described above, NLS is still waiting on the implementing regulations for this section; however, we would like to consider how the potential for expanded linkage to administrative data may be accommodated in our framework.

Taken as a whole, these legal requirements support the adoption of a framework that:

1. requires information that uniquely identifies respondents to be suppressed from the PUF;
2. applies additional safeguards to data elements for which the risk of harm from compromised privacy is elevated; and
3. considers the utility of the elements to data users.

Confidence of Data Providers

The second broad consideration in our framework comprises alternate perspectives from which we think about promoting and preserving the confidence of survey respondents and other data providers. The perspectives focus on tending the public trust; maintaining strong relationships with data providers; and limiting potential privacy harms.

1. Tending the Public Trust

The perspective of tending the public trust motivates a set of objectives that reinforce the dual mandate of the legal framework discussed above. Fundamentally, promoting the public trust entails performing two basic functions: a) continuously and carefully identifying and protecting against anticipated privacy threats; and b) practically and equitably maximizing the public benefit of research that uses the data.

a. Protecting against anticipated threats

The first imperative for tending the public trust is to protect against threats that could be reasonably expected. Many of the tasks that NLS already performs would fall into this category; it entails review of each variable to identify whether a) the data element itself is prone to a risk of being used for identification; and b) there are any values that could potentially identify a respondent. Detailed geographies belong in the former category, and the program should evaluate new variables for similar concerns as it develops its questionnaires. The second category may include a wide range of elements, including categorical variables with many possible values, such as college majors, and continuous variables that may include extreme values, such as income and assets.

An emerging source of threats that the public would expect the survey to be aware of is the ability to link records to publicly available data. The program should continually scan the environment to maintain awareness of any such sources. It should also consult with the privacy community to stay aware of tangible threats that may emerge.

b. Maximizing the public benefit

An equally high priority for tending the public trust is to ensure that the collected data may be put to effective use. This imperative inhabits NLS's whole data production process, affecting design of the collection instrument, data collection, processing of collected data, and provisioning of data access. The

program should refer to the FCSM Data Quality Framework to maximize data quality along its many dimensions – of which confidentiality is one. For the purpose of allocating data elements among access tiers, the program should analyze projected research uses of the data to assess the potential effects on research of applying SDL treatments to data on the PUF and/or moving the element to a RUF. Questions to consider include: a) how frequently is the element expected to be used? b) is the element necessary for exploratory research such as may be appropriate for the PUF? c) are fine levels of accuracy needed to support the research that would likely use the element? For data elements that have appeared in previous NLS surveys, these questions may be considered in light of past uses; for new elements, the analysis may look to the experiences of other surveys or the program's own justification for collecting the data.

2. *Maintaining strong relationships with data providers*

A second approach for promoting the confidence of data providers is to view the task through the lens of NLS's ongoing relationships with the providers. The basic intuition of this approach is that the program should take deliberate steps to honor its promises to data subjects and providers.

The first component of this process is to ensure that the treatment of collected survey data is consistent with the consent statements agreed to by the respondents. As noted above, this is already a regular practice of the NLS program. In addition, the program should pay attention to any additional expectations of privacy protection that the respondents may have. For example, when NLS collects data through confidentiality-preserving tools such as self-administered questionnaires, that may confer an expectation that the data will be disseminated with additional protections.

As NLS considers expanding its use of alternative data sources, such as through linkages to administrative records, it will also have to incorporate expectations and express requirements that accompany the linkages. If consent is given to link records, does it come with additional restrictions? If linkages are performed based on implied consent, are any restrictions also implied? If the alternative data are provided by another Federal agency or other source, does that provider have requirements for the restriction of access? NLS should confront each of these questions as it evaluated the composition of its PUF and RUFs.

3. *Limiting potential privacy harms*

A third approach to safeguarding the confidence of data providers is to focus on harm containment. This paradigm is consistent with the Evidence Act's mandate to assess the sensitivity of each data asset. As described by Ohm (2015), the key questions that the program should consider in assessing the sensitivity of each data element are:

   a. *Can the element be used (by adversaries) to cause harm?* Here we should look particularly to potential material harms that could be suffered if a data breach occurred. Information about criminal involvement is one example of a sensitive element – knowledge by others could have any number of negative ramifications. Information about geographic locations is another exemplar, as it could be used by adversaries to target the subject for attack.

b. *Is there a sufficiently high probability of harmful application?* The second component to consider is whether the potential harm has a high enough likelihood to warrant concern. A potential harm that has an extremely remote chance of occurring should be weighed deliberately against the public benefits of user access. As noted by the CEP, this component and the first are the primary factors for which assessment is needed to manage risk.
c. *Have any special warranties been made about avoiding this harm?* Ohm's paradigm calls for the consideration of "modern harms," which may include psychic effects such as impacts on dignity or one's sense of autonomy. Similar to its analysis through the relationship lens, NLS could operationalize its consideration of such harms by asking whether the data provider may have special expectations of privacy around particular data elements.
d. *Is the analysis overly affected by a majoritarian perspective?* Finally, Ohm includes a check on whether a special need may exist to protect the privacy of a particular sub-group. For example, if particular answers to a question might carry risk of privacy harm, the program should consider that possibility.

Application of Framework to NLSY data – an initial investigation

To explore how the framework above may be applied to the NLSY26 and anticipate some sense of the allocation outcomes that will be supported, we apply the framework to the data elements in a few domains, using NLS's historical experiences in these areas as a guide. Three domains are discussed: geography, health, and crime/justice.

1. *Geography*

The NLS database includes varying levels of detail on the residential geography of its respondents; these variables have been very important for researchers to have in performing a wide variety of studies. In addition to direct and primary uses of the geographical data (e.g., for understanding the effects of place on individuals), geographic information provides useful context for many others and is often identification in studies on a wide range of topics (e.g., for understanding the effects of laws and policies that may differ by location).

As it is traditionally seen as an important determinant of re-identification risk in microdata sets, the level of geographical detail is a primary driver of how previous NLS cohorts have delineated their tiers of data access. At present, the geographic variables on the PUF that describe where the respondent lives at the date of the interview consist of (1) a categorical variable for region of the country with 4 values (Northeast, South, Midwest, and West), (2) an indicator variables describing the urbanicity of the respondent's residence, and (3) a categorical variable indicating whether the respondent lives in a core-based statistical area (CBSA). Two RUFs contain more detailed geographic information. First, state and county FIPS codes are available on the geocode file (VDE File). Second, zip codes and FIPS census tract codes are available at the BLS National Office and at FSRDCs (Zip code file). These nested layers of detail allow for exploratory and summary-level work to be performed using the PUF while enabling qualified researchers with a specific need to access more detailed information.

As the previous experience of the NLS program has suggested that they strike a good balance between utility and confidentiality protection, we expect that application of the framework will recommend that geography be handled in a similar manner in the new cohort. While both RUFs contain geographic units that could be tied to a small population, any output from the VDE and Zip code files must undergo disclosure review before being released to the researcher. This system has been well understood and accepted by data users and respondents, and past examinations have suggested the risk of reidentification is low. Aside from the concomitant harms that arise from reidentification, the potential harm to respondents of having their location revealed at the census tract or zip code level are moderate.

Note that, because they are needed for contacting respondents, the residential addresses of NLS respondents is known, even if it is not shared in any datasets. It may be possible to use this very detailed information to create useful ecological measures. Currently, one summary ecological variable, local unemployment rate in the month of the respondent's interview[4], is available on the geocode restricted use file and available through the NLS VDE. Potentially, a wide variety of ecological variables could support research using NLSY data, such as matches of the address to measures of particulate air pollution or house value. Data elements like these could be very valuable to researchers but would have to be considered very carefully for reidentification risk. NLS could explore offering such data on a restricted basis by offering special-use RUFs; this would need to be supported by formal processes and IT resources to ensure the preservation of confidentiality. An alternative may be to create ecological variables that are fuzzed enough to protect confidentiality while retaining analytical utility. In general, NLS does not expect to release variables based on geography as part of the PUF that permit the identification of geography at or below the state-level (including categorical variables) since the accumulation of geographic-based variables could lead to re-identification of the geographic unit.

2. Health

For previous cohorts, the NLS program has released all health information collected from respondents on the PUF. In the NLSY97, the health data encompasses items on pregnancy, including whether each pregnancy ended in a live birth, miscarriage, or abortion; use of substances such as cigarettes, alcohol, marijuana, and additional drugs; mental health such as number of days of work or school missed in the last 12 months because of emotional, mental, or psychiatric problems, and item-level responses for anxiety and depression indices; and reports on a range of health conditions that limit employment or schooling by type of condition ranging from learning disabilities to blindness or deafness to chronic conditions. While we view none of these items as revelatory, we will consider respondents' comfort with such information being publicly released when designating the variables for the RUF.

The health data in the NLSY has been used extensively and productively by social scientists in a range of fields including economics, sociology, psychology, and public health. NLS hopes that constructing variables that provide summary information on sensitive health items and making them available on the PUF might be adequate for some researchers, while restricting access to more detailed information on these topics. For instance, for the PUF, NLS could consider releasing variables on substance use that indicate whether a respondent used a particular substance within the specified time period, not at all, once, or more than once, while releasing the detailed responses about how much and when a

[4] This variable is reported for metropolitan and micropolitan statistical areas in which the respondent lives. If the respondent resides outside of a metropolitan or micropolitan statistical area, the value is for the portion of the state that is not part of a metropolitan or micropolitan statistical area.

respondent drank alcohol or used drugs on the RUF. Similarly, variables indicating the specific conditions or illness that a respondent has might be released on a RUF, with summary variable indicating that the respondent has a health condition that limits work or school available on the PUF.

3. Crime and Justice

A valuable contribution of the NLSY97 was its collection of information about criminal activity and interactions with the justice system. To the best of our knowledge, the NLSY97 is the only data source with event history data on arrests, charges, sentencing, and incarceration along with detailed individual characteristics, and a myriad of outcomes measured over the life cycle. All of data on these topics were released on the PUF. The NLSY97 crime and justice data have been used in over 200 journal articles and at least 50 dissertations.

At least two administrative databases exist and could be linked to NLSY97: (1) Criminal Justice Administrative Records System (CJARS), longitudinal records of criminal justice proceedings across jurisdictions from 20 states and (2) the Jail Data Initiative (JDI) generated from webscraping the daily jail rosters for over 1,300 counties. Neither of these data sources includes juvenile records. Both require permission to access and link the individual records.

CJARS resides in the FSRDCs and matching would require approval from both Census and BLS, with disclosure review through Census. We know less about the process of gaining access to the JDI individual records from the NYU Public Safety Lab. In theory, the JDI individual records could be linked to NLSY data on incarceration using month and year of birth, sex, and race/ethnicity, and being incarcerated on a given date. Confidence of such a match would be questionable, but if accurate could identify an NLSY respondent given that JDI individual records contain name.

Because of the approval required to obtain these administrative databases, neither pose threat of re-identification, but their recent arrival and that JDI is constructed by web scraping public records suggests that NLS will need to reconsider releasing justice information on the PUF for the NLSY26 and may necessitate moving the NLSY97 justice information for years 2019 and later to the RUF.

Conclusions

Several lessons are emerging from this ongoing work. First and foremost, both sides of the framework – the application of existing and emerging legal requirements and the practical focus on maintaining the confidence of data providers – highlight the importance of considering the research value of the data as well as the possibility of privacy harm. A focus should be on ensuring that practical hazards are anticipated and addressed in a way that does not significantly limit the research value of the data. Second, our initial investigations suggests that past and current NLS practices are broadly consistent with the optimal future approach; we would expect the NLSY26 to employ a tiered-access approach similar to that currently in place for the NLSY79 and NLSY97. However, NLS should add additional components to its assessments of suitability for inclusion in each tier of data access. It should carefully consider the sensitivity of particular data elements that it collects, with an emphasis on the potential for use of the information to cause material harm. Even if there is no immediate hazard identified and the

likelihood of harm is low, prudence counsels that sensitive data elements should be afforded additional safeguards. Following this effort, we expect that a substantial number of variables that are on the public-use file (PUF) in the NLSY79 and NLSY97 will move to a restricted-use file (RUF) in the NLSY26.

Third, the NLS program will have to give consideration to a number of additional issues that may emerge as new data elements are considered for inclusion in the NLSY26. It will need to give special attention to how linkages to other confidentiality are handled, as well as data that may be created by linkage to public sources on the basis of detailed geography. Throughout this work, the program will have to carefully consider the resource needs of accommodating its tiered access users.

Bibliography

Advisory Committee on Data for Evidence Building: Year 2 Report, October 14, 2022. https://www.bea.gov/system/files/2022-10/acdeb-year-2-report.pdf

Federal Committee on Statistical Methodology. "A Framework for Data Quality." (September 2020). https://www.fcsm.gov/assets/files/docs/FCSM.20.04_A_Framework_for_Data_Quality.pdf

Federal Committee on Statistical Methodology. "Statistical Policy Working Paper 22: Report on Statistical Disclosure Limitation Methodology." (2005). https://www.fcsm.gov/assets/files/docs/spwp22WithFrontNote.pdf

Milner, Justin, Katharine Abraham, Ron Haskins, Karen Pittman, and Kathy Stack. "Realizing the Promise of Evidence-Based Policymaking." https://www2.census.gov/adrm/fesac/2017-12-15/Abraham-CEP-final-report.pdf

Ohm, Paul. "Sensitive Information." Southern California Law Review, vol. 88, no. 5, July 2015, pp. 1125-1196.

Statistical Policy Directive No. 1: Fundamental Responsibilities of Federal Statistical Agencies and Recognized Statistical Units" (*79 FR 71609*).

Statistical Policy Directive No. 4: "Release and Dissemination of Statistical Products Produced by Federal Statistical Agencies" (*72 FR 42266*).

Table 1. Variable Categories in the NLSY97 Restricted Geocode File

Location	Education	Survey and Created Variables
State/county/MSA of residence	College UNITID	State born
Primary Sampling Unit	College State	country born
state parents born	A little bit from transcripts	state child live
continuous unemployment rate	Major field of study	round 7 timings
residence info. Age 12	College application info.-college IDS	names of state welfare program
region for parents and grandparents	pending admission decision	information about welfare participation
maternal/paternal grandparents-same country	pending financial aid decision	time limits for programs-welfare-medicaid-food stamps
	term for application	other information for programs
	GED State	state grandparents born
		military veterans-medals

Distance	County and City Data
Migration	Characteristics Merged in from County-City Data Book
state/county/quality/impute	
Various migration measures-flags	
geocode distance to mom, dad	

Table 2. Variable Categories in the NLSY97 Restricted Zipcode and Census Tract File

Location	Education	Migration
zip code	Secondary school identifiers	Zip code
census tract		Indicators for whether state/county/zip imputed
country parents, grandparents born		

Table 3. Examples of Variables in the Restricted Geocode File Vs. Public Use File

Geocode	Public Use
State of residence at interview date	Census region (4) at interview date
College UNITID	ID that allows researcher to trace whether specific respondent attends same college across rounds
Geocode distance to mom, dad	Collapsed distance to mom, dad
Military veterans-medals	Collapsed categories of medals