Best Practices in Administrative Data Collection that Facilitate Research: Lessons Learned from the NCSP Exploratory Data Analytics Study

August 30, 2024

Authors

Albert Y. Liu, Jenny Nguyen

Submitted to

U.S. Department of Labor
Chief Evaluation Office
200 Constitution Ave NW
Washington, DC 20210

Project Officer

Janet Javar

Submitted by

Westat, Inc.
1600 Research Boulevard
Rockville, MD 20850

Project Director

Allison Hyra

Disclaimer

This report was prepared for the U.S. Department of Labor (DOL), Chief Evaluation Office (CEO), by Insight Policy Research (now Westat) under contract number 1605DC-18-A-0018. The views expressed are those of the authors and should not be attributed to DOL, nor does mention of trade names, commercial products, or organizations imply endorsement of same by the U.S. Government.

Acknowledgments

The authors are grateful to Tara Martin at the U.S. Department of Labor for her championship and guidance on this project. We would like to thank many people for their excellent input, review, and feedback, including staff from the CEO and OSHA at the Department of Labor, as well as the Evaluation Support team at Avar Consulting.

Suggested Citation

Liu, A. Y. & Nguyen, J. (2024). *Best practices in administrative data collection that facilitate research: Lessons learned from the NCSP Exploratory Data Analytics Study*. Westat, Inc. U.S. Department of Labor Chief Evaluation Office.

Contents

What's Inside This Brief?

What's here? This report provides recommendations for implementing administrative data collection practices that support the research needs of Federal departments and agencies. We first describe how we assessed the feasibility of using administrative data from the U.S. Department of Labor's (DOL) Occupational Safety and Health Administration (OSHA) to conduct an implementation study of the National Construction Safety and Health Achievement Recognition Program Pilot (NCSP). We then present recommendations for enhancing administrative data collection practices to better advance evidence-building and expand evaluation capacity.

Why focus on improving administrative data collection procedures? The Foundations for Evidence-Based Policymaking Act of 2018 (Evidence Act) requires federal departments and agencies to produce Learning Agendas and Annual Evaluation Plans that describe their efforts to use evidence in decision-making and build evidence where it is lacking. Federal departments and agencies already collect a great deal of administrative data, but their data collection practices are not always systematic and structured. When administrative data is collected in these ways, departments and agencies will need to complete labor-intensive data processing, such as classifying the data and completing data entry, before using it for evaluation. Enhancing existing administrative data collection practices can serve as a cost-effective way for departments and agencies to comply with the Evidence Act, including strengthening the evaluation capabilities of Federal staff.

Who should read this report? We designed this report to support program staff in Federal departments and agencies who collect and manage administrative data. Recognizing administrative data collection can vary greatly in who provides the data, how the data is provided, and what the data is used for, we provide recommendations that Federal departments and agencies can tailor to meet their context and goals. Our recommendations are motivated by programs that serve individuals and/or employers, although they can also apply to Federal activities such as enforcement and grant monitoring. State and local agencies, as well as any organizations collecting administrative data, can also benefit from the recommendations outlined in this report.

How do I use this report? We recommend readers begin with the introduction to understand the purpose of the report and then move to the specific application of using administrative data for research in the NCSP Exploratory Data Analytics Study. Readers can then review the administrative data practices that facilitate research and some additional resources for advancing evidence-building and expanding evaluation capacity.

Section I. Introduction

Federal departments and agencies are required to advance evidence-building and expand evaluation capacity. The Foundations for Evidence-Based Policymaking Act of 2018 (Evidence Act) (Pub. L. No. 115-435) requires Federal departments and agencies to create Learning Agendas that describe how they will use data to answer important short- and long-term strategic and operational questions. In particular, it requires them to address "challenges to developing evidence to support policymaking, including any statutory or other restrictions to accessing relevant data." Relatedly, in the Office of Management and Budget (OMB) memo M-21-27 about the Evidence Act, OMB urged departments and agencies to use "existing evidence, sometimes in new ways and contexts… to support processes like agency operations, grantmaking, human capital management and development, and program administration, as well as to support mission strategic areas, like program and service delivery" (OMB, 2021).

Federal departments and agencies already collect a great deal of administrative data, defined as information about individuals or programs collected and maintained as part of program operations (Kline, 2022). However, their data collection practices are not always systematic and structured. When administrative data is collected in these ways, departments and agencies will need to complete labor-intensive data processing, such as classifying the data and completing data entry, as a first step before using it for evaluation. Compared to new data collections, modifying existing administrative data collection practices can be a cost-effective way for Federal departments and agencies to better use their data (Kline, 2022).

The U.S. Department of Labor (DOL) has been at the forefront of efforts to use administrative data to conduct research (DOL, 2024). In addition to completing the requirements set forth by the Evidence Act (see box 1) and conducting its own primary data collection efforts for research and evaluation studies, DOL's Chief Evaluation Office (CEO) holds Administrative Data Research and Analysis (ADRA) contracts. These contracts allow CEO to conduct ad hoc, quick turnaround studies that leverage DOL's administrative data and provide timely responses to inform strategic agency priorities.

This brief describes ways Federal departments and agencies can enhance their administrative data collection practices to facilitate research based on lessons learned from the ADRA NCSP Exploratory Data Analytics Study. CEO recently initiated this study to examine the feasibility of using administrative data from DOL's Occupational Safety and Health Administration (OSHA) to conduct an implementation study of the National Construction Safety and Health Achievement Recognition Program (SHARP) Pilot (NCSP). CEO provided the study team from Westat Insight with internal NCSP administrative data. The study team then reviewed the data, developed a process to turn the data into machine-readable data files that can be processed by a computer, and completed a preliminary implementation study. The process of turning the data into research-ready files was labor intensive because the NCSP administrative data in PDF and JPG forms required manual data entry to convert the text into machine-readable data. Federal departments and agencies interested in using administrative records to support research can apply lessons learned from this study to enhance their own practices and evidence capacity.

Box 1: U.S. Department of Labor Implementation of the Evidence Act

- **Chief Evaluation Officer.** The Chief Evaluation Officer leads CEO to coordinate, manage, and implement the DOL evaluation research program. This person is the Evidence Official who oversees implementation of Title I of the Evidence Act, including guiding the Evaluation Plan, Learning Agenda, and Evidence Capacity Assessment.
- **Chief Data Officer.** The Chief Data Officer leads implementation of Title 2 of the Evidence Act. This person is responsible for data governance and lifecycle data management and leads efforts to improve data management processes, ensure that data are fit for purpose, and increase capacity to use data as a strategic asset.
- **Annual Evaluation Plan.** CEO develops an Evaluation Plan (e.g., https://www.dol.gov/sites/dolgov/files/OASP/evaluation/pdf/DOL-CEO-FY-2024-2025-Evaluation-Plan.pdf) each year to describe the evaluation activities it plans to undertake to address areas of strategic importance. The plan provides a list of each project, along with its priority research question, the data to be used to answer that question, and the method for conducting the study.
- **Learning Agenda.** Every 4 years, DOL develops its Learning Agenda (https://www.dol.gov/sites/dolgov/files/evidence/evidence-building-plan-fy2022-2026.pdf), which describes the department's activities to generate information about important short- and long-term strategic and operational questions.
- **Evidence Capacity Assessment.** The DOL Capacity Assessment (https://www.dol.gov/sites/dolgov/files/evidence/evidence-capacity-assessment.pdf) describes the current state of the coverage, data quality, evidence-building methods, effectiveness, and independence of its statistics, evaluation, research, and analysis activities. The report describes areas of strength and weakness in the use of evidence by staff and the use of evidence in decision-making.

The remainder of this brief provides an overview of the NCSP Exploratory Data Analytics Study (section II), describes administrative data collection practices that support research (section III), discusses the conclusions (section IV), and lists additional resources to support the collection of high-quality administrative data (section V).

Section II. The NCSP Exploratory Data Analytics Study

NCSP is administered by OSHA through State On-Site Consultation programs[1,2] participating in the pilot. NCSP allows OSHA On-Site Consultation programs to work with small- and medium-sized[3] businesses in the construction industry to develop and implement comprehensive and effective safety and health management programs for all workers on-site. Highly trained occupational safety and health professionals (i.e., consultants) from On-Site Consultation programs provide no-cost on-site consultation services to employers that voluntarily request such services. Employers must meet specific occupational safety and health requirements to be approved to participate in the construction SHARP Pilot.

Historically, participation in SHARP was only available to single fixed worksites. However, small businesses in the construction industry experience the highest rates of fatal and non-fatal injuries. The construction industry had the second-highest number of fatal injuries in 2022 (U.S. Bureau of Labor Statistics, 2023), with small businesses accounting for over half (57 percent) of all fatal injuries from 2011 through 2022 (Harris et al., 2024). OSHA intends to use NCSP administrative data to determine if the construction SHARP Pilot should be included in the general criteria for SHARP participation. To this end, OSHA reached out to the DOL's CEO to request assistance with assessing the NCSP data collection process and determining how NCSP data could be effectively utilized.

OSHA first developed a research agenda with definitions of measures of implementation fidelity, along with outcomes and impacts (OSHA, 2022), and its plans were included in the DOL Fiscal Years 2023–2024 Evaluation Plan (DOL, 2023). CEO employed its ADRA contractor, Westat Insight, in 2022 to transform the internal NCSP administrative data from 6 states covering data from 2015 to the present into machine-readable databases and determine what they say about the NCSP implementation fidelity.[4]

The study team conducted a deep dive to understand the NCSP program and administrative data. The team conducted a document review of program manuals and developed an NCSP logic model that included program intake and consulting services for SHARP certification and renewal as activities and documentation of identified hazards and remedies and self-assessment reports as outputs. The study team also discussed research priorities with OSHA to determine the most important data elements relevant for an implementation study.

As the first step of the implementation study, the study team studied and processed two types of NCSP administrative data. The first was an extract from the OSHA Information System (OIS) database, which consultants and other operational staff use to process and store consultation data, including data on

[1]The OSHA On-Site Consultation Program offers no-cost and confidential occupational safety and health services to small- and medium-sized businesses in all 50 states, the District of Columbia, and several U.S. territories, with priority given to high-hazard worksites. Consultants from State On-Site Consultation programs located within local agencies or state universities, work with employers to identify workplace hazards and how to fix them, provide advice for compliance with OSHA standards, train and educate, and assist in establishing and improving safety and health programs (learn more at osha.gov/consultation).

[2] Two NCSP participating Consultation programs were not included in the administrative data study because they were newly approved participants without construction projects for the study.

[3] Typically, participation in NCSP is available only to employers with 250 or fewer employees onsite and no more than 500 employees corporation wide. The upper corporate size limit does not apply to individual franchisees.

[4] The data are internal to OSHA and not publicly available.

construction projects approved to participate in the SHARP Pilot. The second was administrative documents saved for each construction project by On-Site Consultation programs. Based on the study team's review, the documents were either standalone file types or combinations of file types from the following list:

- **Employer's Request.** The Request Form—known as Form 20—contains information about each consultation visit request made by the employer. The form documents the request number that uniquely identifies each construction project, employer's name, worksite address, training and assistance services requested, and points of emphasis noted by a consultant.
- **Consultation Visit.** Known as Form 30, the Consultation Visit form contains data about each consultation visit and is completed by a consultant. The form documents the scope of the visit, which can be safety, health, or both; whether the visit was full-service or limited-service; information about the training provided to the employer; and the points of emphasis noted by a consultant.
- **Case Status.** The Case Status Form summarizes every consultant visit for a construction project under a specific Employer's Request number. The form catalogs every visit number; whether each consultation type was an initial, follow-up, or training and education visit; whether the service requested was for safety, health, or both; visit dates; and the date the written report describing the visit findings was delivered to the employer.
- **Letter.** The Letter record type consists of formal correspondence. Examples of letters include cover letters of written reports by consultants, recommendations for SHARP approval by consultant project managers (CPMs), and cover letters for completed Employer Reports of Action Taken by employers. They do not include emails unless they were used as the cover letter for a written report.
- **Written Report Body.** The Written Report Body record type documents the findings of a consultation visit. It typically has an executive summary specific to the construction project that documents the employer's North American Industry Classification System (NAICS) code and the Total Recordable Case Rate (TRC) and Days Away, Restricted, and/or Transferred (DART) rates for the project, industry in the State, and industry in the country. It also typically contains programmatic language that applies across all the NCSP projects, such as sections on Employer's Obligations and Rights, Interim Protection for Employees, Evaluation of Safety and Health Management System, and Other Findings and Recommendations. The Written Report Body record type is used for the body of a written report to the employer, and the Letter record type is used for the cover letter of a written report.
- **SHP Assessment Worksheet.** The Safety and Health Program (SHP) Assessment Worksheet, known as Form 33, is used by a consultant to assess the safety and health program at a worksite (see figure 1). Based on the safety and/or health visit findings, the consultant rates the worksite on seven SHP elements: hazard anticipation and detection, hazard prevention and control, planning and evaluation, administration and supervision, safety and health training, management leadership, and employee participation. Scores for each of the 58 assessment attributes on the form range from 0, indicating the required safety or health procedure or policy is not present, to 3, indicating the safety or health procedure or policy is exemplary. Eight of the assessment attributes are marked as stretch items for the employer (i.e., safety and health attributes beyond the basic attributes of a safety and health program). The SHP Assessment Worksheet is either stored as a stand-alone file or included as an appendix to a written report.

Figure 1. Example of the Safety and Health Program Assessment Worksheet Form 33

Safety and Health Program Assessment Worksheet
Form 33

Request Number		**Visit Number**	**Visit Date**
Employer			
Site Location			
Legend: 0=No; 1=No, Needs major improvement; 2=Yes, Needs minor improvement; 3=Yes; NA= Not Applicable; NE= Not Evaluated; *=Stretch items Attribute of Excellence			

Synthesis Item Score	**Score**
Hazard Anticipation and Detection Score	
Hazard Prevention and Control Score	
Planning and Evaluation Score	
Administration and Supervision Score	
Safety and Health Training Score	
Management Leadership Score	
Employee Participation Score	
Total Score	
Average Score	

- **List of Hazards**. The List of Hazards record type is typically included as an appendix to a written report and is to be posted unedited in a prominent place where it is readily observable by all affected employees for 3 working days, or until the hazards are corrected, whichever is later. The file documents the hazards found during the consultation visit, hazard types, number of instances, standards violated, correction due date, and conditions and descriptions of the hazards.
- **Report of Hazards Found.** This record type contains a comprehensive list of hazards the consultant identified during a safety and/or health consultation visit. For each hazard, the consultant records the hazard type, number of instances, standard violated, correction due date, hazard description, and recommended corrective action for the employer. Reports of Hazards Found are often, but not always, included as an appendix to a written report.
- **Hazard Summary.** The Hazard Summary record type stores information about every hazard found during a consultation visit. For each hazard, the form documents the item number, number of instances, standard violated, hazard type, number of workers at risk, correction due date, date the hazard was corrected, and verification date and method the consultant used to verify that the employer corrected the hazard.
- **Employer Report of Action Taken.** This record type can be provided by a consultant to the employer to complete and is submitted by the employer to the consultant after completion. Employer Report of Action Taken files consist of a list of the hazards found during a consultation

visit, hazard types, numbers of instances, correction due dates, dates hazards were corrected by the employer, standards violated, corrective actions taken, and actions taken to prevent reoccurrence. The files are often included as an appendix to a written report.

- **Action Plan.** The Action Plan describes the steps that the employer will take to ensure continuous improvement of its worksite safety and health management system. The Action Plan includes the construction project and company TRC and DART rates and the national average of the TRC and DART rates for employers in the same industry (i.e., with the same NAICS code).
- **Work-Related Injuries and Illnesses.** The OSHA Work-Related Injuries and Illnesses form is a two-part form employers use to maintain occupational injury and illness records. The first part is Form 300, which an employer completes to log every work-related death, injury, or illness that involved the loss of consciousness, a restricted work activity or job transfer, and days away from work in a calendar year. The second part is Form 300A, which is a summary of Form 300.
- **Other Documents.** The last form type is Other Documents, which is applied to all other types of administrative records. Other Documents include Consultation Activity Sheets used by some States, surveys and questionnaires developed by CPMs for participating employers, industrial hygiene monitoring results in written reports, consultant field notes, emails, and employer safety manuals.

The study team learned that many of these file types were based on Word templates created using the OIS data structure and stored by On-Site Consultation programs in the OIS. As such, the documents of a single file type had the benefit of having a roughly similar structure in terms of table layouts, keywords, and contents. However, the Word templates were not locked, so consultants were able to adjust the structure to suit their needs. Many but not all of the files were saved in construction project folders, and individual documents didn't always indicate the project to which they belonged. To prepare the files for data processing, the study team manually separated combined files into files of a single file type and tagged each one with the project number. About one-third of the files were document scans in PDF format, and because optical character recognition software designed to read and process text contained in images was unable to do so effectively and consistently, the study team completed manual data entry to obtain data from those files.

The study team then developed a data-scraping procedure to record relevant information from each file. Data scraping involves pulling information out of a file and placing it into a spreadsheet (Mitchell, 2018). The study team developed and ran a series of Python scripts, one per file type, to scrape the files, and used the resulting machine-readable files. OSHA ultimately provided six batches of data as it located them, and the study team completed this process for each one except the final batch, which OSHA provided when the initial analysis was already complete. The study team then used the machine-readable data files to conduct an implementation study based on measures that were available in the NCSP administrative data.

The implementation study confirmed that the NCSP administrative data could be used for research, but there were several challenges to overcome. These challenges include:

- **The NCSP administrative files could be combinations of multiple file types.** The study team conducted a manual review process to separate and tag individual files. The study team processed 773 administrative files for the data scraping procedure and created 1,402 individual files that were each a single file type.

- **Not all files were in a machine-readable format.** The study team needed to complete manual data entry for files that were not in a machine-readable format. Many files were poor quality scans and thus did not contain machine-readable text after using optical character recognition.[5] Manually entering the data for these files was labor intensive.
- **Consultants had different recordkeeping practices.** The study team found that there were many instances of preliminary drafts of files that were saved alongside final drafts of files. OSHA also continued to locate additional files as the data review progressed.
- **There were systematic differences in recordkeeping policies.** The study team also found that states had differing data retention policies. One state, in particular, instituted a policy of only retaining files for 5 years. As a result, the study team was unable to access those files, which limited the meaning of the implementation findings for that state.

To support OSHA in building its research agenda, the study team developed a series of data collection recommendations. Those recommendations form the basis of the following recommendations in section III for Federal departments and agencies more broadly.

[5] The study team experimented with using Adobe Acrobat, Microsoft Word, and Python for optical character recognition.

Section III. Administrative Data Practices that Facilitate Research

We used the lessons learned throughout the NCSP Exploratory Data Analytics Study to develop recommendations that can apply to Federal departments and agencies to support compliance with the Evidence Act (see table 1). The first decision-making stage focuses on determining what data elements to collect as administrative data. The second stage addresses best practices for collecting and storing administrative data.

Table 1. Summary of recommendations by decision stage

Stage	Recommendations
1. Determine the data elements to collect as administrative data	1. Develop a comprehensive understanding of the program 2. Develop the research questions 3. Determine what data elements are most important to include in the administrative data and their sources
2. Efficiently collect and store the required data elements	1. Use data collection strategies that yield machine-readable data 2. Use data collection forms with consistent file layouts 3. Use data collection strategies that yield high-quality data 4. Define the required recordkeeping procedures

1. Determine the data elements to collect as administrative data

Recommendation 1. Develop a comprehensive understanding of the program

As a first step, it can be helpful for Federal departments and agencies to thoroughly understand the program environment that generates the administrative data. Federal departments and agencies can complete this process by creating a program logic model, which illustrates how a program works and how it achieves its goals (W.K. Kellogg Foundation, 2010). A logic model such as the one in table 2 illustrates the following components of a Federal agency program, although not all components will apply to some programs:

- **Inputs.** Program inputs are the available community, human, financial, and organizational resources used.
- **Activities.** Activities are the processes, tools, and actions that are part of program implementation.
- **Outputs.** Outputs are the direct results of activities delivered by the program, such as the number of processes, tools, and actions provided.
- **Outcomes.** Outcomes are the changes in participant behavior, knowledge, skills, status, and functioning.

- **Impacts.** Impacts are the long-run changes that occur in communities and organizations.
- **Contextual factors.** These are factors outside of the program that affect how the program operates and the outcomes of its participants.

Table 2. Example of a program logic model

Inputs	Activities	Outputs	Outcomes	Impacts
Workforce investment system	Screening and assessments	Basic skills	Placement in unsubsidized employment	Increased wages
Education and training	Job readiness and soft skill development	Vocational skills	H1-B employment	Hours (full-time/part-time)
Business-related nonprofits	Vocational training	Credentials	Increased wages	Benefits
Employers	Supportive services	Work-based learning	Hours (full-time/part-time)	Increased career mobility
Child care	Job development and placement services	Soft skills	Benefits	Household well-being

Source: Adapted from Gasper et al., 2021.

Logic models can be applied not only to programs but also to enforcement activities undertaken by Federal departments and agencies. For example, the DOL Employee Benefits Security Administration (EBSA) carries out Title I of the Employee Retirement Income Security Act (ERISA). EBSA could develop a logic model for handling civil and criminal investigations, with the investigation components listed as activities and employer compliance with ERISA as the outcome.

Federal departments and agencies can develop a logic model through a variety of means. There are several resources available that they can use as guides, such as El Mallah and colleagues (2022). They can also speak directly with program officers and front-line staff to understand their roles and responsibilities. Federal departments and agencies may conduct document reviews, such as program manuals that describe how a program is supposed to operate, as well as administrative data already being collected to see how the program operates in practice. In the case of the NCSP Exploratory Data Analytics Study, the team reviewed the SHARP policies and procedures manual that defines the program activities, reviewed the NCSP administrative data, and then developed a draft logic model and revised it to incorporate OSHA feedback (OSHA, 2023).

Recommendation 2. Develop the research questions

Once there is a logic model that describes the program, Federal departments and agencies can define any research questions that they wish to have answered to comply with the Evidence Act requirement on submitting a plan that includes questions for developing evidence to support policymaking. Typical research

questions often ask about implementation, outcomes, or impact, with example questions presented in table 3. Implementation studies address the extent to which programs operate as intended, and outcomes studies address whether programs achieve the immediate outcomes that are closely tied to the programs (OMB, 2020). Impact evaluations are designed to assess the causal effect of an intervention (OMB, 2020).

Table 3. Potential research questions for the NCSP by study type

Study type	Sample questions
Implementation Study	• What activities are consultants implementing with 100 percent fidelity? • What activities are not being implemented with 100 percent fidelity, and how do the rates differ between states with federally versus state-administered programs?
Outcomes Study	• How many construction projects participated in the NCSP and were approved for SHARP? • How do participant counts differ between states with federally versus state-administered programs?
Impact Study	• What is the impact of the NCSP on safety and health practices? • How do the impacts differ between states with federally versus state-administered programs?

Defining research questions can help Federal departments and agencies focus their attention on where to build evidence to support decision-making. In the case of the NCSP Exploratory Data Analytics Study, refining the research questions allowed OSHA to classify their research questions as implementation, outcome, or impact and understand what types of study designs are needed for each group.

Federal departments and agencies can develop research questions by conferring with program officers and the agency Chief Evaluation Officer. They can also solicit ideas from stakeholders and program participants. Federal agencies may also consult the research base to determine which research questions have already been answered. El Mallah and colleagues (2022) can serve as a guide for developing research questions and, ultimately, evaluation designs. Once the research questions have been defined, they can be considered for inclusion in the Evaluation Plan and Learning Agenda.

Recommendation 3. Determine what data elements are most important to include in the administrative data and their sources

Once the logic model is in place and there are defined research questions, Federal departments and agencies can define the set of required data elements and their sources to comply with the Evidence Act requirement on submitting a plan that describes the data they intend to collect and analyze to support the use of evidence in policymaking:

- **Define the data elements.** The first step is to define the data elements needed for each logic model element or to answer a research question. The logic model data

elements are likely to be implementation outputs and participant outcomes. For example, in an employment and training environment, implementation outputs might be the number of training sessions and other services provided and participant outcomes might be employment status and wages. For research questions, the elements might be contextual factors that are relevant for subgroup analyses, such as geographic location, industry, or participant demographic characteristics.

- **Determine the source for each data element.** The second step is to determine the best source for each data element. Some data elements can be provided by multiple people: for example, implementation outputs can be provided either by implementation staff or by program participants. Other data elements can be provided as an automatic part of program operations, such as output from a case management system.

It is important to identify the data elements and their sources to ensure that the administrative data contains the required information to answer the research questions. When crucial data elements for answering research questions are missing, departments and agencies will be unable to answer their research questions unless they expend resources on new data collections.

Federal departments and agencies may want to identify a comprehensive set of ideal data sources when thinking through what is best answered with administrative data. In the case of the NCSP Exploratory Data Analytics Study, the study team mapped out ideal data sources for implementation, outcome, and impact research questions that included not only administrative data, but also employer surveys, surveys of consultants, and interviews and focus groups with CPMs. The team limited its attention to the administrative data elements for the implementation feasibility study, although OSHA may consider other data sources for future research activities.

2. Efficiently collect and store the required data elements

Recommendation 4. Use data collection strategies that yield machine-readable data

Federal departments and agencies can consider modifying their administrative data collection practices to ensure that they yield machine-readable data, which is data that can be read and processed by a computer without human intervention. Administrative data stored in databases such as management information systems and files where individual text can be selected, such as Word documents, are machine-readable.

Collecting administrative data that is not in machine-readable form can result in labor-intensive data processing. For the NCSP Exploratory Data Analytics Study, PDF and JPG images were particularly problematic because they required manual data entry to convert the text into machine-readable data. Furthermore, due to limited study resources, the team did not convert any of the images of safety and health management system violations into data, even though the data in them could be useful for some research activities.

Federal departments and agencies seeking support in these efforts may want to consult with their Chief Data Officer and IT support staff. Identifying administrative data can be challenging. For example, some PDF documents support the ability to yield machine-readable data, while others do not. Converting PDF forms to ones that yield machine-readable data can be completed by sufficiently skilled IT support staff combined with clear instructions for the individuals filling out the forms.

Recommendation 5. Use data collection forms with consistent file layouts

Federal departments and agencies can consider modifying their administrative data collection practices so that human input is based on a consistent and fixed form. Examples of fixed forms are databases with manual data entry forms, Word documents that do not allow users to modify table layouts and field labels, and PDF survey forms. Such forms will generate administrative documents with a consistent structure that facilitates exporting or file scraping for future research.

The challenge with forms that can be modified—if they are modified—is that they increase the level of effort needed to export the data into a machine-readable database. Scraping processes using rules-based automation procedures can handle expected changes to a file layout but not unexpected changes, such as the addition of tables or rows within tables. In the case of the NCSP Exploratory Data Analytics Study, the team realized that some SHP Assessment Worksheets had a single date while others had multiple dates and were able to write Python code to accommodate both types of records. However, unexpected changes to the file layout, such as additional rows in the Visitation Worksheet, required manual review and coding adjustments to scrape the intended fields.

Recommendation 6. Use data collection strategies that yield high-quality data

Administrative data collection practices can be designed to yield high-quality data. Federal departments and agencies can consider using data validation procedures. Data validation procedures are processes that ensure valid values for the most important data elements, such as a check that a staff member assigned to the participant is a real employee, a year is a feasible year, or that a participant's state of residence is actually a state. Federal departments and agencies can also consider using administrative data collection practices that minimize differences between data providers. That is, they can be structured so that data provided by different people in the same role provide the same data elements. For example, data on program activities can be designed to come consistently from either program managers or front-line staff but not an unspecified mixture of the two.

The availability of high-quality administrative data is crucial for enabling high-quality research based on that data. Missing data can limit the usefulness of research findings and can require labor-intensive human intervention to impute when it is possible to do so. In the case of the NCSP Exploratory Data Analytics Study, the study team was able to impute the construction project number associated with each file based on a

review of the folder structure, the file contents, or a question to OSHA staff, but that tagging process was very labor intensive.

Recommendation 7. Define the required recordkeeping procedures

The final recommendation is for Federal departments and agencies to have clear and binding recordkeeping procedures. Staff contributing to administrative data should have clear instructions about what data to keep, where to store the data, and how long to store the data. The instructions should also be designed to support the data collection required to meet the research needs of the Federal department or agency. Federal departments and agencies should also define recordkeeping procedures when data can change over time, such as retaining all versions of the data but archiving previous versions. To the extent possible, administrative data should not be overwritten as new information arises, but instead retained for later use to allow departments and agencies the ability to see how that data element has changed over time.

When instructions about recordkeeping are unclear or not aligned with the research needs, the available data can either require labor-intensive imputation approaches or limit the data availability for the research study. For the NCSP Exploratory Data Analytics Study, the study team learned that one state had a data retention policy of only 5 years: all administrative data older than 5 years was to be deleted. Because the study took place more than 5 years after some of the construction projects, it was limited in that it could not differentiate between missing files and unretained files for those projects.

Section IV. Conclusion

Federal departments and agencies have latitude in how they collect administrative data, and how they do so has implications for how well the data meet their needs. Federal departments and agencies collect administrative data for a variety of purposes, including as part of regular program operations and, more recently, to support research to support the implementation of evidence-based policies and programs as per the Evidence Act.

This brief drew on lessons learned from the ADRA NCSP Exploratory Data Analytics Study. The purpose of the study was to transform the NCSP administrative data into machine-readable databases and determine what they say about the implementation fidelity to assist OSHA with its goal of determining how NCSP data could be effectively collected and utilized. The study team conducted a deep dive to understand the NCSP program and administrative data and developed a data-scraping procedure to record relevant information from each file. The study team faced some substantial challenges in completing this work, including variations in file structure and recordkeeping practices and the use of files that were not in a machine-readable format.

This brief uses those challenges identified in the NCSP Exploratory Data Analytics Study to provide potential solutions and best practices that Federal departments and agencies might consider when determining how to collect administrative data as required by the Evidence Act. At a high level, the best practices consist of recommendations for determining which data elements to prioritize and how to collect those data elements in an efficient way to help Federal departments and agencies collect administrative data that can meet their various needs with minimal additional resources. Implementing these practices can help Federal departments and agencies efficiently enhance their own data collection practices and evidence capacity as required by the Evidence Act.

Section V. Additional Resources

- DOL Evaluation Plan: https://www.dol.gov/sites/dolgov/files/OASP/evaluation/pdf/DOL-CEO-FY-2024-2025-Evaluation-Plan.pdf
- DOL Evidence Building Plan: https://www.dol.gov/sites/dolgov/files/evidence/evidence-building-plan-fy2022-2026.pdf
- DOL Capacity Assessment: https://www.dol.gov/sites/dolgov/files/evidence/evidence-capacity-assessment.pdf
- Federal guidance on learning agendas: https://www.whitehouse.gov/wp-content/uploads/2019/07/M-19-23.pdf
- Federal guidance on evaluation standards: https://www.whitehouse.gov/wp-content/uploads/2020/03/M-20-12.pdf
- Logic model development: https://www.acf.hhs.gov/sites/default/files/documents/prep-logic-model-ts_0.pdf
- Research question development: https://www.acf.hhs.gov/opre/toolkit/program-managers-guide-evaluation

Appendix. Full Bibliography

El Mallah, S., Gutuskey, L., Hyra, A., Hare, A., Holzwart, R., & Steigelman, C. (2022). *The program manager's guide to evaluation* (OPRE Report 2022-208). U.S. Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research, and Evaluation. https://www.acf.hhs.gov/opre/toolkit/program-managers-guide-evaluation.

Foundations for Evidence-Based Policymaking Act of 2018. Pub. L. No. 115-435, 132 Stat. 5529. (2019). https://www.congress.gov/115/plaws/publ435/PLAW-115publ435.pdf.

Gasper, J., Gearing, M., Giesen, L. Marrow, J., Muz, B., & Dodkowitz, A. (2021). *Evaluation of the TechHire and Strengthening Working Families Initiative Grant Programs: Findings from the implementation study.* U.S. Department of Labor. https://www.dol.gov/sites/dolgov/files/OASP/evaluation/pdf/TechHire-Implementation-Report-v2.pdf.

Government Performance and Results Act of 1993. Pub. L. No. 103-62, 107 Stat. 285. (1993, August 3). https://www.congress.gov/bill/103rd-congress/senate-bill/20/text.

Harris, W., Brooks, R. D., Trueblood, A. B., Yohannes, T., & Bunting, J. (2024). *Fatal and nonfatal falls in the U.S. construction industry, 2011-2022*. Data Bulletin: March 2024. https://www.cpwr.com/wp-content/uploads/DataBulletin-March2024.pdf.

Kline, N. (2022). *Using administrative data in social policy research* (OPRE Report 2022-163). U.S. Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research, and Evaluation. https://www.acf.hhs.gov/sites/default/files/documents/opre/administrative_data_brief_feb2023.pdf.

Mitchell, R. (2018). *Web scraping with Python*, 2nd Edition. O'Reilly Media.

Occupational Safety and Health Administration. (2023). *Consultation policies and procedures manual*. https://www.osha.gov/sites/default/files/enforcement/directives/CSP_02-00-005.pdf.

Occupational Safety and Health Administration. (2022). *On-site consultation program National Construction: SHARP Pilot Program (NCSP) annual evaluation: Goals, strategies, and measures*.

Office of Management and Budget, Executive Office of the President. (2020). OMB Memorandum M-20-12. *Phase 4 implementation of the foundations for Evidence-Based Policymaking Act of 2018: Program evaluation standards and practices.* https://www.whitehouse.gov/wp-content/uploads/2020/03/M-20-12.pdf.

Office of Management and Budget, Executive Office of the President (2021). OMB Memorandum M-21-27. *Evidence-based policymaking: Learning agendas and annual evaluation plans*. https://www.whitehouse.gov/wp-content/uploads/2021/06/M-21-27.pdf.

U.S. Bureau of Labor Statistics. (2023). *Census of fatal occupational injuries summary, 2022* (USDL-23-2615). https://www.bls.gov/news.release/cfoi.nr0.htm.

U.S. Department of Labor. (2023). *U.S. Department of Labor evaluation plan: Fiscal years 2023 and 2024*. https://www.dol.gov/sites/dolgov/files/evidence/DOL-CEO-FY-2023-2024-Evaluation-Plan.pdf.

U.S. Department of Labor. (2023). *U.S. Department of Labor evaluation plan: Fiscal years 2024 and 2025*. https://www.dol.gov/sites/dolgov/files/OASP/evaluation/pdf/DOL-CEO-FY-2024-2025-Evaluation-Plan.pdf.

W.K. Kellogg Foundation. (2010). *W.K. Kellogg Foundation logic model development guide*. https://www.naccho.org/uploads/downloadable-resources/Programs/Public-Health-Infrastructure/KelloggLogicModelGuide_161122_162808.pdf.