Attachment I – Project Topic

Fraud in Public Healthcare Programs: Developing Fraud Detection Models with Linked Data

Key Objectives

Fraud detection across public programs is hampered by critical data being scattered among disparate systems, formats and agencies (see 2024 GAO report). This fragmentation creates significant information gaps that hinder the government's ability to identify which new and expanding programs are most vulnerable to fraud. This challenge is compounded by limited data sharing and interoperability across oversight bodies, making it difficult to assemble a comprehensive view of fraud. However, by standardizing and linking unstructured, open-sourced, and under-utilized data, agencies can detect emerging patterns of fraud more quickly and build timely evidence.

This Request for Solutions (RFS) seeks to show how existing federal and state databases could be used to produce new data products that augment understanding of fraud trends and to create new risk-scoring tools that can be leveraged by the Department of Justice's (DOJ) Fraud Section's Healthcare Fraud Unit and others. The objectives of this project are to demonstrate the feasibility and value of using the National Secure Data Service (NSDS) to combine data from federal, state, and other sources to: (a) develop a new risk-scoring model to detect the magnitude of fraud against federally-funded public healthcare programs; (b) identify risk factors and vulnerabilities for healthcare fraud; and (c) pilot the use of new technologies that can lend efficiencies to currently underutilized or previously unlinked data sources.

Background

The National Secure Data Service

The CHIPS and Science Act, Section 13075(c), calls for engagement with federal and state agencies through an NSDS demonstration project to "collect, acquire, analyze, report, and disseminate statistical data in the United States and other nations to support governmentwide evidence-building activities consistent with the Foundations for Evidence-Based Policymaking Act of 2018."

The National Center for Science and Engineering Statistics (NCSES) at the U.S. National Science Foundation is a federal statistical agency with authority in law and expertise in practice to protect Americans' confidential data. The CHIPS and Science Act positioned the NSDS demonstration project at NCSES with clear mandates about privacy protections. It specifically called for NSDS to protect "confidential data and statistical products" and to ensure that "no individual entity's data or information is revealed by the National Secure Data Service demonstration project platform to any other party in an identifiable form." Any datasets resulting from linkages with restricted data have the same protection as the restricted data. Linking restricted data with a public dataset, for example, results in a linked dataset that is itself restricted.

All NSDS projects using restricted data must be performing statistical activities for statistical purposes. These terms are defined in the Confidential Information Protection and Statistical Efficiency Act (CIPSEA) as follows:

> ‘‘(10) STATISTICAL ACTIVITIES.—The term ‘statistical activities’— ‘‘(A) means the collection, compilation, processing, or analysis of data for the purpose of describing or making estimates concerning the whole, or relevant groups or components within, the economy, society, or the natural environment; and ‘‘(B) includes the development of methods or resources that support those activities, such as measurement methods, models, statistical classifications, or sampling frames.
>
> ‘‘(12) STATISTICAL PURPOSE.—The term ‘statistical purpose’— ‘‘(A) means the description, estimation, or analysis of the characteristics of groups, without identifying the individuals or organizations that comprise such groups; and (B) includes the development, implementation, or maintenance of methods, technical or administrative procedures, or information resources that support the purposes described in subparagraph (A).

A Data-Driven Approach to Combatting Healthcare Fraud

In a recent report, the US Government Accountability Office (GAO) estimated that the federal government lost between $233 billion and $521 billion to fraud each year between 2018 and 2022. The same report noted there are substantial data-related challenges to developing a comprehensive picture of fraud against federal programs. It recommended a more systematic approach that better leverages data to understand this pervasive and costly problem.

Additionally, in March 2025, the White House released an Executive Order aimed at “Eliminating Information Silos,” which directs agencies to take steps to ensure that “officials designated by the President or Agency Heads (or their designees) have full and prompt access to all unclassified agency records, data, software systems, and information technology systems...for purposes of pursuing Administration priorities related to the identification and elimination of waste, fraud, and abuse."

The Health Care Fraud Unit has a Memorandum of Understanding (MOU) with the Centers for Medicare & Medicaid Services (CMS) to access Medicare Part B claims in part to routinely analyze healthcare fraud trends over time. Using these data, the Healthcare Fraud Unit’s Data Analytics Team has returned results in identifying healthcare fraud trends that far exceed the costs of staff, equipment and data sources. It has been successful in identifying geographic hotspots for fraud and other data outliers.

Going forward, increased cross-agency linkage of data could support a more comprehensive assessment of the potential risk factors for various types of fraud, and such an assessment could benefit a broad array of federal programs. Thus, the opportunity to utilize NSDS' capabilities for both expanded data access and an environment for data linkage would allow the Healthcare Fraud Unit to provide a statistical and evidence-based look at the magnitude of undetected fraud as an alternative to the traditional enforcement approach, which historically has not relied on the use of statistical and evidence-based analysis. Combining the Health Care Fraud Unit’s subject matter expertise on indicators and behaviors with robust research and statistical data analysis within a sharable cloud environment could serve as a model to encourage substantially greater data sharing among agencies.

The Current Request for Solutions

This project will involve the following key activities:

1. Reviewing the literature and documentation on health care fraud to identify existing fraud taxonomies, definitions, and measures, indicators and behaviors.
2. Modifying the existing MOU with CMS for Medicare Part B claims data, unless the project activities are determined to fall within the scope of the existing MOU.
3. Identifying and documenting primary sources of complaints and allegations of health care fraud.
4. Identifying and documenting privacy, legal, security, and institutional barriers to accessing federal and state healthcare data.
5. Obtaining, if feasible given project time constraints, an MOU for data sharing with at least one state-level medical claims data source.
6. Developing standard MOU specifications to address requirements. Primary sources should include the Department of Health and Human Services (HHS) Office of Inspector General (OIG) Hotline data and Medicare Part B claims data.
7. Conducting webscraping on open-source platforms for verified patient reviews (including, but not limited to, ZocDoc and similar platforms) to identify and collect additional data to inform the fraud model.
8. Using the NSDS Secure Compute Environment, develop one or more fraud detection models that leverage linked data, using Artificial Intelligence (AI), Machine Learning (ML), and other emerging technologies.

It will be essential for this project to have an agile and collaborative project team, including key individuals from the Healthcare Fraud Unit and the DOJ Criminal Division's Office of Policy and Legislation (OPL); identified representatives and signatory officials with significant legal and data acumen from data source offices; and state complaint board representatives. It will also require frequent and transparent collaboration and communication between DOJ technical experts and the project performer to incorporate subject matter expertise on fraud indicators. Upon award, the awardee should create and maintain an up-to-date list of project team members and their contact information throughout the duration of the project.

Information Gaps

This project will support:

- The feasibility of using the NSDS to explore key questions for large-scale federal programs and will demonstrate the use of NSDS to develop models to assess fraud in public programs.
- Examples of success in multi-agency coordination and thoughtful data linkage to limit the project's need for novel data or development of new MOUs.
- The application of explainable AI, to help with data standardization and to efficiently prepare unstructured data for use in the NSDS PPRL environment.
- Usage of the NSDS Secure Compute Environment in the development of a fraud risk-scoring algorithm for health care fraud.
- The future development and application of a validation approach for health care fraud.

Key Evidence Building Considerations

- What is the magnitude of undetected fraud in select public healthcare programs?
- What are the patterns of fraud in select public healthcare programs over time?
- What are program characteristics, indicators, or risk factors associated with fraud against select public healthcare programs?
- Are there program characteristics, indicators, or risk factors associated with fraud in these select programs that might be usefully adapted to identifying and measuring fraud against other public programs?

Deliverables

At a minimum, the following deliverables will be provided to the participating agencies:

- Monthly status reports on progress towards project objectives.
- Quarterly lessons learned based on what has been learned during the last quarter that will inform a future NSDS.
- Draft MOUs to serve as templates for additional similar efforts or continuations of the project.
- A briefing paper detailing the process used to mine and standardize unstructured data from HHS, state complaint boards, and other services such as ZocDoc, Inc.
- An analysis plan specifying methods to be used for: cleaning the data, treatment of missing data, linking/merging data, etc. which should be delivered for review prior to analysis.
- A PowerBI (or similar, open-source tool) dashboard to allow for querying and visualizing outliers, with the potential to link future data sources. It is expected that the dashboard should be appropriate for both technical and non-technical audiences and will conform to the NSDS purpose of statistical activities for statistical purposes described above.
- All analytical outputs (e.g. a linked dataset, codebook, algorithms) at the project's close.
- All code and documentation supporting the project suitable for inclusion in a publicly accessible open-source software repository and provided under an open-license or public-domain equivalent; including all code used in the project provided in a reproducible open-source format (R, python, etc.). All technical components should propose using open-source software and open-source software development principles wherever feasible.
- Data (and/or documentation of data) sufficient to reproduce the results of all statistical analysis including those done using open-source software
- A statistical assessment using the cross-agency data assessing the statistical trends and risk factors associated with fraud against federal programs.
- A briefing paper on the risk-scoring algorithm.
- A final report covering the project objectives and outcomes that is to a standard acceptable for public dissemination by a federal agency.