Office of the
CHIEF TECHNOLOGY OFFICER

THE STATE OF DATA SHARING AT THE U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES

September 2018

The Data Initiative Team
Office of the Chief Technology Officer
U.S. Department of Health and Human Services

ACKNOWLEDGEMENTS

The Office of the Chief Technology Officer in the Department of Health and Human Services would like to thank the many contributors who made this report possible. Individuals from the following agencies generously gave their time and shared their vast knowledge with us: Administration for Children and Families, Administration for Community Living, Agency for Healthcare Research and Quality, Centers for Disease Control and Prevention, Centers for Medicare and Medicaid Services, Food and Drug Administration, Health Resources and Services Administration, National Center for Health Statistics, National Institutes of Health, and Substance Abuse and Mental Health Services Administration.

This report was written by the Data Initiative Team at the HHS Office of the Chief Technology Officer with support from staff from the Administration for Children and Families' Office of Trafficking in Persons, the Centers for Disease Control and Prevention's National Center for Health Statistics, and the National Institute of Health's National Institute of Allergy and Infectious Diseases.

TABLE OF CONTENTS

Introduction 4
Methods 6
Challenges 8
Process for Data Access 8
Technology for Data Access & Analysis 11
Regulatory Environment 14
Disclosure Risk Management 20
Norms & Resource Constraints 23
Next Steps 26
Appendix 27
List of Agencies & Data Assets 27
List of Acronyms 33

INTRODUCTION

The potential for machine learning, artificial intelligence, and other tools to address some of the most complex challenges, to promote discovery, and to augment the possibilities of applied human intelligence has become increasingly accepted.[1] And yet, the difficulty of the journey of creating value from data is underestimated when the sole focus is on an ultimate output rather than the foundational elements necessary to embed analytics as a core function of organizational performance. As Fortune 500 companies are recognizing this imperative of analytics, their primary driver is revenue. Federal and state governments are on a similar transformational journey of using data to advance mission delivery. The move to make more data publicly available, under the banner of the open data movement, is seen as essential to making government itself more open. Public access to open data also enables data consumers including entrepreneurs, innovators, and researchers to use data to generate new products and services, build businesses, and create jobs. Indeed, the cultural shift of closed to open government data has resulted in more than two thousand datasets published to HealthData.gov for the public to discover and use.[2] While the value proposition for open data has taken root in the marketplace, government agencies must likewise use its data as a strategic asset.

Across the twenty-nine distinct agencies of the United States Department of Health and Human Services (HHS), data essential to understanding the nation's health are collected every day. [3,4] Whether surveillance, survey, or claims data, HHS expends an enormous amount of financial resources to report on the state of the health of the population it serves. These data, however, are largely kept in silos with a lack of organizational awareness of what data are collected across the Department and how to request access. Each agency operates within its own statutory authority and each dataset can be governed by a particular set of regulations. As such, each discrete analysis of the data often gets reviewed for legal purposes and leads to data sharing occurring largely on a project-by-project basis. The individuals involved negotiate the nature and extent of data sharing arrangements often

[1] Bughin, J., Hazan, E., Chui, M., Allas, T., Dahlstrom, P., Henke, N., & Trench, M. (2017, June). McKinsey Global Institute: Artificial Intelligence, The Next Digital Frontier. Retrieved from https://www.mckinsey.com/~/media/McKinsey/Industries/Advanced Electronics/Our Insights/How artificial intelligence can deliver real value to companies/MGI-Artificial-Intelligence-Discussion-paper.ashx

[2] HealthData.gov. (n.d.). Retrieved September 11, 2018, from https://healthdata.gov/search/type/dataset

[3] HHS agencies are also referred to as Operating Divisions and Staff Divisions.

[4] HHS.gov. (2015, October 27). HHS Agencies & Offices. Retrieved September 11, 2018, from https://www.hhs.gov/about/agencies/hhs-agencies-and-offices/index.html

based on past experience and personal relationships. The process can lack transparency, transferability, accountability, and consistency.

Data governance is defined as a set of processes created to ensure that data assets are formally managed throughout the enterprise. A data governance model establishes authority, management, and decision-making parameters related to the data produced or managed by the enterprise.[5] Across the federal government, there is growing consensus in the value and promise of data governance to reduce inefficiencies and costs. These goals align with the Bipartisan Commission on Evidence-based Policymaking Report[6]; Cross-Agency Priority (CAP) Goal 2: Leveraging Data as a Strategic Asset of the President's Management Agenda (PMA)[7]; the HHS Digital Government Strategy: Building a 21st Century Platform to Better Serve the American People[8]; and multiple objectives of HHS Strategic Goal 5: Promote Effective and Efficient Management and Stewardship.[9]

A cohesive enterprise-wide data governance strategy that promotes data sharing, drives business value from leveraging data as an asset, and bases policies on evidence is essential to a long-term data-driven vision of HHS.

[5] NIST Computer Security Resource Center. (n.d.). Retrieved September 11, 2018, from https://csrc.nist.gov/Glossary/?term=3846

[6] US CEP. (2017, September 29). CEP Final Report: The Promise of Evidence-Based Policymaking. Retrieved September 11, 2018, from https://www.cep.gov/cep-final-report.html

[7] General Services Administration, The Office of Management and Budget. (2018, June 26). President's Management Agenda: Modernizing Government for the 21st Century. Retrieved from https://www.performance.gov/PMA/Presidents_Management_Agenda.pdf

[8] HHS.gov. (2018, June 28). Digital Strategy at HHS. Retrieved September 11, 2018, from https://www.hhs.gov/web/governance/digital-strategy/index.html

[9] HHS.gov. (2018, February 28). Strategic Plan FY 2018 - 2022. Retrieved September 11, 2018, from https://www.hhs.gov/about/strategic-plan/index.html

METHODS

This report focuses specifically on data assets identified by the agencies as having high value and categorized as restricted or nonpublic. In the context of this report, restricted data generally refers to data, which contain personally identifiable information and therefore cannot be shared publicly. Nonpublic data, again in the context of this report, refers to programmatic or administrative data, such as grant application submissions, which are not gathered for the purpose of sharing but may have valuable secondary uses. Data collected for statistical purposes may also be restricted or nonpublic and are available in Research Data Centers or other protected environments.[10] To learn more about these datasets, over the course of the past several months, we conducted semi-structured interviews with leadership and agency personnel from eleven HHS agencies. Our objective was to understand the challenges and opportunities to the sharing of restricted and nonpublic data among HHS agencies.

Initially, we met with senior leaders from the different agencies to understand their priorities, concerns, and evolving strategies regarding their internal data portfolios. We also gathered information and perspectives about high-level data sharing and governance policies. Further, we sought recommendations for data assets at their agency to conduct follow up interviews. Factors that determined data asset selection included importance within the agency, perceived value to other agencies, and the extent of challenges faced when sharing that data.

With these recommendations, we selected twenty-seven data assets about which to conduct detailed semi-structured interviews (See Appendix). Each interview lasted up to 2 hours. During these semi-structured interviews, staff most familiar with the data assets were asked to discuss the characteristics and conditions influencing the sharing of that data.

Questions were guided by domain areas developed by the team. Domain areas included secondary use cases, descriptions of the data sharing process; data security, quality, timeliness, and lifecycle; metadata management; confidentiality and privacy concerns; technical approach to sharing; and, any analytic support offered. In addition to learning about the data assets, the team also identified some data-related best practices and challenges as defined by the agencies.

[10] Title V of the E-Government Act, Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA) (Pub. L. 107-347, title V; 116 Stat. 2962, Dec 17, 2002) defines statistical purpose as "the description, estimation, or analysis of the characteristics of groups, without identifying the individuals or organizations that comprise such groups; and includes the development, implementation, or maintenance of methods, technical or administrative procedures, or information resources that support the purposes" described as statistical activities. (https://www.gpo.gov/fdsys/pkg/PLAW-107publ347/html/PLAW-107publ347.htm).

Detailed notes were taken by multiple team members during each interview. Within three days of an interview, working notes were created and circulated among the team for verification, correction, and elaboration as needed. The team also developed frameworks for organizing semi-structured interviews that allowed insights from the interviews to emerge. Following final acceptance of the interview notes, each data asset and its program were reviewed to characterize the findings related to data sharing policies and practices, opportunities, and limitations.

The findings in this summary report are based on many hours of interviews and follow up review. We believe these generalizations to be true and reflective of the current state of data sharing among the agencies, but we also recognize that they do not apply to all data systems and that this report is limited by our contact with the selected data asset stewards. We did not speak with all staff from all data assets from across the Department. Nevertheless, during this process, we identified core challenges that inhibit the sharing of restricted and nonpublic data among HHS agencies and the following section of the report will delineate these findings. Addressing these challenges will enable HHS to begin its journey in becoming a data-driven organization that better serves the American People.

Challenge 1
Process for Data Access

HHS lacks consistent and standardized processes for one agency to request data from another agency. Agencies are not accountable for their responses to requests for access to internal data. If access is inappropriately denied or if access is significantly and inappropriately delayed, there are no consequences.

The Department lacks a consistent, transparent, and standardized framework for sharing restricted and nonpublic data among its agencies in a timely and efficient manner. Each agency, and often agency personnel for each dataset, has the autonomy to interpret the rules for data sharing processes. Data sharing processes can range from non-existent and informal, to formal and consistent such as those present at the Research Data Centers (RDCs).[11] While datasets shared through the RDC have well-defined procedures available online, for many datasets, the data governance rules are not formalized. The sharing of those datasets can be ruled by individual relationships and/or staff availability. This appears to be more frequently the case for datasets that are shared rarely or datasets that are relatively new.

Typically, data are shared among agencies by using various forms to document the exchange and the data requestor's acceptance of any restrictions on the data. Data sharing documents range widely from agency to agency. Some agencies use Memorandum of Understanding (MOUs), Data Sharing Agreements (DSAs), Data Use Agreements (DUAs), or Interagency Agreements (IAAs). The variation exists not only from one agency to another but within agencies as well. Once a data analyst requests access to data using one of these agreements, the language in these documents can be reviewed and revised several times over a period of several months to a year before the data analyst can gain access. This lengthy process creates a burden on the data analyst to navigate a range of agreements to access datasets from within HHS. For a data analyst new to HHS, this presents a significant barrier to entry in using HHS data as an asset. Most agency personnel interviewed referred to the multiple efforts required by the requestor to complete the agreements and described the length of the process as highly variable, depending upon the complexity of the request and staff availability.

In addition to variability in accessing and using various forms, there is great variability in how data sharing is governed at different agencies. Many agencies reported an absence of a data governance group at the agency level. Thus, there is often no systematic approach for tracking all requests and the outcomes of those requests. One data representative shared, "We track requests through an email inbox. [The primary point of contact] who oversees the email will answer and respond on his own. If he is unable to respond, the request is forwarded to [other staff]". The lack of an agency or department-wide

[11] The National Center for Health Statistics developed the Research Data Center to protect the confidentiality of survey participants, both individuals and institutions, while allowing researchers access to these restricted data. The NCHS RDC hosts restricted data from various groups within HHS. Access is granted to researchers who have proposals approved by the RDC. Criteria for approval include the demonstrated need for access to sensitive data, appropriately planned analyses, limitations of potential disclosure risk, and other characteristics (https://www.cdc.gov/rdc/index.htm). Many agencies make their data available through the Census Research Data Center, which are scattered throughout the U.S. There are currently 28 open Federal Statistical Research Data Center (RDC) locations (https://www.census.gov/about/adrm/fsrdc/locations.html).

governance structure not only leads to difficulty in navigating various procedures and forms but can also lead to a lack of accountability regarding access requests. These processes lead to variability in agency response. There was little evidence of agency-level procedures to address concerns if access is significantly and inappropriately delayed or denied altogether. Currently, there are no consequences for inappropriate delay or denial of data sharing.

The lack of standardization at the departmental level for data governance and sharing, the lack of accountability for timely response to requests, and the fact that data are largely kept in silos, often results in HHS agencies having no means to access interagency data in an efficient way.

Challenge 2
Technology for Data Access & Analysis

The technical formats and approaches to sharing restricted and nonpublic data across agencies vary widely. The analytical tools to interpret data can be redundant. Finally, agencies are tracking who has access to restricted and nonpublic data but can be challenged in auditing analyses for misinterpretation and misuse.

Machine-readable file formats are the majority but not yet widespread default

The majority of data programs offer to share data in multiple machine-readable formats; however, one-fourth of the data programs share data in one file format, some of which are PDFs. Nearly half the programs offer multiple static file formats (Excel, CSV, SAS, etc.) when sharing data; five programs offer additional analytic or visualization tools in addition to static formats; and two data assets offer application programming interfaces (APIs) as a data access method. The remainder of programs share data in one file format. While not pervasive, this presents a challenge for data analysts because it is not a given that data will be shared in a machine-readable format, creating an additional labor-intensive step. Yet, select programs, such as the National Adult Maltreatment Reporting System from the Administration for Community Living, are creating an API for data ingestion and sharing.

There is opportunity to collaborate in software and data acquisitions across agencies

Once received, data analysts across agencies employ a variety of technologies to ingest, cleanse, and analyze the data. There is a decentralized approach to selecting, acquiring, and managing these applications. While not the focus of this report, there are significant redundancies in the instances of technologies across the Department. Efforts are ongoing to understand these redundancies in the landscape of goods and services purchased across HHS' agencies. The CDC's Data Hub Program presents a promising solution to acquiring data at an agency-wide level to reduce costs and increase efficiencies.[12] Greater transparency is required in the acquisition of both software and data to eliminate inefficiencies and one-off project based uses in service of a Department-wide data and software acquisition strategy.

Preventing data misinterpretation remains a strong priority but contingent on resources

A point of interest and tension is the degree to which analytic support and auditing can take place after the data are shared. Agency representatives for both survey data and administrative data sources are not always assured that their data will be handled and contextualized properly once it is shared. As one agency personnel said, "Misuse of the data is an issue. We monitor this on a monthly basis." There is reluctance and apprehension about whether proper stewardship and interpretation of the data will be employed after data are shared. For agencies sharing sensitive data, this is magnified

[12] U.S. Department of Health and Human Services, Centers for Disease Control and Prevention. (2018, June 19). CDC Data Hub. Retrieved from https://www.cdc.gov/ophss/csels/dhis/documents/dhis-data-hub-508.pdf

further. As one agency said, “Once the [entity] gives us their data, they lose all control. We are stewards of their data,” implying the high level of scrutiny applied in considering requests for restricted and nonpublic data.

All agency personnel stated that they track who has access to restricted or nonpublic data. However, the manner in which data are tracked is inconsistent and, at times, difficult. Agencies whose data collection are governed under the Confidential Information Protection and Statistical Efficiency Act (CIPSEA) use a Research Data Center (RDC) to share restricted or nonpublic data. At the RDC, individuals must come to a facility to access the data, and the agency reviews and evaluates the analyses before an individual leaves the RDC. This guards against misapplication, inappropriate disclosures, or data leaks. The Administration for Children and Families uses a database to track access to National Directory of New Hires. CMS also maintains a system called the Enterprise Privacy Protection Engine (EPPE) to track the release of individually identifiable disclosures.

To track what analyses are conducted with restricted or nonpublic data, some agency personnel provide technical assistance, assess analyses before public release, or retrospectively audit the analyses for misinterpretation of the data. The degree to which agency personnel can audit the analysis of data and provide technical support is often limited. When asked about auditing and technical support for data analysts, the agency, with a robust contract in place, said, “[Our contractor] has subject matter experts who can help with these kinds of issues.” Another agency said, “We provide all the support that we can give to our users, but we cannot fund additional staff.” A select few agencies had contract resources to search and audit publications citing agency data. Agencies expressed the importance of the needed capability to enhance the auditing and tracking of data assets and associated analyses, particularly of restricted and nonpublic data.

Challenge 3
Regulatory Environment

Each data collection effort has statutes, regulations, and policies that govern the collection of and access to the data. Some statutes limit access to data and its use. In order to increase access or broaden use, changes to the relevant statutes may be required.

During each interview with leadership and dataset representatives, we discussed the different laws and regulations, as well as the varying interpretations of the statutes authorizing collection of and governing access to restricted and otherwise nonpublic data at HHS.

Statutes are often the foundation authorizing the collection of the data. In addition to statutes authorizing the collection of information, there are separate statutes pertaining to privacy and confidentiality protections. Further, there are some agency-specific legislative statutes related to data use and confidentiality protections. In some instances, statutes clearly specify the groups of people who can access the data and the purposes to which the data may be applied. Some agencies will not release restricted or nonpublic data unless statutory authority is explicitly specified. The National Directory of New Hires (NDNH) under the Administration for Children and Families is an example of such a data collection effort. The NDNH, the national database of wage and employment information, is used to assist states administering programs that improve their abilities to locate parents, establish paternity, and collect child support. Personal Responsibility and Work Opportunity Reconciliation Act of 1996 (PRWORA), which established the NDNH database, represents a statute with clearly defined provisions governing access to the data. The NDNH requires that a person or group be specifically authorized through the statute in order for the data to be shared. The Administration for Children and Families publishes a guide communicating who can access various types of data and how they may utilize it.[13] In such cases, data sharing beyond what is permitted in the legislation would require a change in statute.

In some instances, a statute can govern data collection efforts across multiple data collection efforts and agencies. Some of these statutes and regulations are described below.

The Privacy Act of 1974

For many of the data collection activities in HHS, the applicable statute governing records about individuals is the Privacy Act. The Privacy Act sets forth a series of requirements governing federal agency practices with respect to certain information about individuals. The law strives to balance the government's need to maintain these records with the individual's right to be protected from unwarranted invasions of personal privacy. The Privacy Act limits agencies to maintaining "only such information about an individual as is relevant and necessary to accomplish a purpose of the agency required to be accomplished by statute or Executive Order of the President." The Privacy Act also requires agencies to

[13] Administration for Children and Families, Office of Child Support Enforcement. (2017, March 10). A Guide to the National Directory of New Hires. Retrieved from https://www.acf.hhs.gov/sites/default/files/programs/css/a_guide_to_the_national_directory_of_new_hires.pdf

"keep an accurate accounting" regarding "each disclosure of a record to any person or to another agency, and to retain the accounting for at least five years or the life of the record, whichever is longer."[14]

The Confidential Information Protection and Statistical Efficiency Act (CIPSEA)

The Confidential Information Protection and Statistical Efficiency Act (CIPSEA) of 2002 (see P.L. 107-347, Title V) is a law establishing confidentiality protections for data collected by U.S. statistical agencies and units. At HHS, there are two entities covered under CIPSEA: The National Center for Health Statistics, the federal health statistical agency, and the Center for Behavioral Health Statistics and Quality, a designated statistical unit. CIPSEA restricts the use of information exclusively to statistical purposes only. Violations of the provisions of CIPSEA are subject to five years imprisonment and/or a fine of up to $250,000. For individuals who request access to restricted data, such as through the Research Data Center, the interpretation of what “statistical purposes” means may seem obscure and the evaluation criteria may be difficult to locate. The language in CIPSEA is very clear and strong about the absolute necessity of maintaining confidentiality; however, this can sometimes lead to difficulty in sharing data for beneficial purposes.[15]

Agency representatives stated that they rely on this legislation to improve their abilities to collect data and to secure the privacy of the information contained in the dataset. Although they recognize the potential for increased data sharing of restricted files, they stressed that access must be accomplished properly and within the regulations created under CIPSEA. Generally, this access is provided through the Research Data Center. Agency representatives were quite clear: “We do not want to ‘loosen’ the restrictions of CIPSEA. We do not see the law as a barrier.”

The Health Insurance Portability and Accountability Act of 1996

Across the Department, agency personnel expressed frustration at the differences in interpretation of the Health Insurance Portability and Accountability Act of 1996 Privacy Rule. The HIPAA Privacy Rule establishes the conditions under which certain individually identifiable health information, referred to as protected health information, may be used or disclosed by covered entities and their business associates, including for research purposes. Covered entities are: health plans; health care clearinghouses; and health care providers who conduct certain financial and administrative transactions electronically.

---

[14] The Privacy Act of 1974, 5 U.S.C. § 552a(c)

[15] Public benefits of activities are defined in OMB’s Statistical Policy Directive No. 1 as having a statistical purpose as (a) relevant and timely, credible and accurate, objective, and protected information informing decision-makers in governments, businesses, institutions, and households (https://www.gpo.gov/fdsys/pkg/FR-2014-12-02/pdf/2014-28326.pdf).

Covered entities and their business associates, with whom they contract with to perform some of their essential functions, are bound by the privacy standards.[16]

Research is defined in the Privacy Rule as, "a systematic investigation, including research development, testing, and evaluation, designed to develop or contribute to generalizable knowledge."[17] Disclosures for public health activities are permitted for certain purposes to a public health authority, which is defined in the Privacy Rule as, "an agency or authority of the United States, a State, a territory, a political subdivision of a State or territory, or an Indian tribe, or a person or entity acting under a grant of authority from or contract with such public agency, including the employees or agents of such public agency or its contractors or persons or entities to whom it has granted authority, that is responsible for public health matters as part of its official mandate."[18]

When negotiating data use agreements or interagency agreements (IAAs), agency personnel expressed that there is a lack of consensus around the interpretation of HIPAA that is very time consuming. Although "research" and "public health" are defined in statute, one agency representative stated, "There is a disconnect with how 'research' and 'public health' are interpreted from one agency to the next [in the HIPAA Privacy Rule]." This causes frustration and dissimilar outcomes across the department.

Title 42 of the Code of Federal Regulations (CFR) Part 2

Title 42 of the Code of Federal Regulations (CFR) Part 2: Confidentiality of Substance Use Disorder Patient Records (Part 2) was intended originally to address concerns about the

social stigma and potential consequences of seeking treatment for a substance use disorder. "Part 2 is intended to ensure that a patient receiving treatment for a [substance use disorder] in a Part 2 program does not face adverse consequences in relation to issues such as criminal and domestic proceedings such as those related to child custody, divorce or employment."[19] The protection 42 CFR Part 2 provides is by restricting access to or disclosure of such treatment records. The ways in which 42 CFR Part 2 has been interpreted, however, have limited the ability of researchers and policymakers to more fully characterize population health issues related to substance use disorders, such as the

[16] HHS.gov. (2002, December 19). Health Information Privacy. Retrieved from https://www.hhs.gov/hipaa/for-professionals/faq/190/who-must-comply-with-hipaa-privacy-standards/index.html

[17] 45 CFR 164.501; HHS.gov. (2018, June 13). HIPAA Privacy Rule: Research. Retrieved from https://www.hhs.gov/hipaa/for-professionals/special-topics/research/index.html

[18] 45 CFR §§ 164.501 and 164.512(b); HHS.gov. (2018, June 13). HIPAA Privacy Rule: Research. Retrieved from https://www.hhs.gov/hipaa/for-professionals/special-topics/research/index.html

[19] The Office of the National Coordinator for Health Information Technology, SAMHSA. (n.d.). Disclosure of Substance Use Disorder Patient Records: Does Part 2 Apply to Me?. Retrieved from https://www.samhsa.gov/sites/default/files/does-part2-apply.pdf

opioid crisis.

State Agreements

Many data collection systems at the Department depend on states voluntarily sharing data. Data timeliness and granularity of the data shared between states and federal agencies can vary greatly. Federal-state legal agreements can exacerbate these challenges. For some data systems, there is no federal requirement for the states to share their data. This can limit the sharing of granular data and the capacity of federal agencies to modify, manage, and improve standard reporting. Several data representatives made it clear during interviews: “We do not own the data,” implying that the states own the data and HHS is a secondary user of the data. When legal agreements are in place, negotiating changes to the data collection of any form would require renegotiating existing agreements and receiving cooperation among state and territory partners. Depending on the data collection system, renegotiation would need to include patients, health systems, and other participants involved with the data collection. When asked about access to data with more granularity than the public file, one data representative said: “Any change to the agreement may have to go to the states, so if [our federal agency] wanted to ‘use’ the data differently, the states and possibly the patients would have to be involved, hence a huge ask,” implying there is significant energy involved in, and thus aversion to, modifying federal-state agreements. Finally, some legal agreements between federal and state partners do not create accountability around data timeliness. Many data systems have strong relationships and management in place for how and when states submit data. However, in some cases, there is a complete absence of any accountability for states to submit data to federal partners in a timely way. Due to several factors, it can take months or years for states to submit. The causes of the delay in state submission are not the focus of this report; however, delayed state data submission has significant implications for data sharing. For internal data sharing, some data representatives did not want to share any provisional data until all states had submitted their data. Data dissemination thus hinges on the last state to submit data, as federal partners are then able to do final cleaning, statistical weighing, and quality checks.

The Interagency Agreement Process

If and when two different agencies share data that is not public, two different legal agreements are typically employed between the agencies. One is an agreement that ensures the data analyst receiving the data are accountable for the use, storage, disclosure, privacy and security, analysis, and dissemination of the data as governed by statutes and policies. The types of agreements employed for this purpose include Data Use Agreements, Memoranda of Understanding, or Data Sharing Agreements.

The second agreement is employed when one agency charges another agency for the data, whether with funds or in-kind contributions. Many agency's general counsels work together to draft and finalize an Interagency Agreement (IAA). Sometimes, the language in the agreement is standardized; however, most agencies, even agencies who regularly charge other agencies for data, describe it as a time-consuming, laborious, and confusing process to coordinate and execute these agreements between general counsels. "The process to finalize an Interagency Agreement [to share data between agencies] is ridiculous," said one agency representative who is frequently engaged in the process. "Fixing that would be a great help to us." It is fair to say that agency personnel are often hesitant to request data when time is of the essence or when there are inadequate resources to manage the IAA process.

Creating these documents and remaining within the law have proven to be a significant challenge across the Department. While no agency disputes the laws and process, the arduous nature of statute, regulation, and policy compliance serves as challenges against widespread data sharing.

Challenge 4
Disclosure Risk Management

The risk of identifying geographic areas or violating individual privacy increases as more variables and more granular data are collected and shared, often leading to an increase in limits on microdata access.

Agency personnel are concerned appropriately about maintaining promised confidentiality and protecting the privacy of citizens' personally identifiable information (PII). However, restrictions related to maintaining this privacy can create substantial challenges to sharing these data, which have the potential for leading to enhanced and integrated programmatic operations. Microdata refers to record level data showing the characteristics of individuals, households, establishments, or other units of analysis.[20] Granularity of data, understood variously as increased specificity of geography or demographic characteristics, finer details of injury or cause of death ICD codes, or more precise racial and ethnic categories, naturally leads to smaller and smaller table cell sizes. In other words, the more microdata and the more granular the microdata, the greater the analytic potential and risk of unintended disclosure.

Personally identifiable information is not collected

In some cases, federal programs collecting administrative data do not collect PII in compliance with the Privacy Act. As one agency personnel stated, "PII is scrubbed by the states before [data] files are sent." In this case, the program requesting the data from the state is careful to not ask for more granular data than they need to perform their programmatic work.[21] In other cases, states remove PII before sending data, which may be required by state regulation. "[We] never get PII or extraneous data elements from states. [As a result] there is no significant difference between the public use file and the restricted use file… The only sensitive field [in the restricted file] is the client's birthday, which is not included in the public file." From another program, an agency personnel noted, "Overall, the critical variables are not more than thirty-five. There are approximately twenty variables in the public use file." The difference in number of variables here is greater, but the intent to protect privacy and confidentiality is the same. By instituting required privacy and confidentiality protections in the data collection methodology, the potential secondary uses of these data become limited.

In cases where the shared dataset does collect and retain PII, restrictions of use and reporting follow the data and are governed by the source of the dataset. One program staff person explained it this way: "[We] must define this data as confidential because the frame for this data are derived from [another agency's] database. We must coordinate the public releases with [the other agency]. [We] cannot release anything that is more detailed than what [the originating agency] releases without [their] consent."

[20] National Center for Health Statistics. (2002, July). Policy on Micro-data Dissemination. Retrieved from https://www.cdc.gov/nchs/data/nchs_microdata_release_policy_4-02a.pdf

[21] The principle of requesting minimal information necessary is pursuant to the Privacy Act of 1974 (5 U.S.C. § 552a(c)).

Laws, consent forms, and governance can limit the amount of microdata shared from national surveys

Many agency staff believe preserving or improving survey response rates depends on strong protections of confidentiality. “Promising confidentiality is not done to limit sharing but to assure that what is shared is worthy of sharing,” one agency representative summarized. In some cases, programs make assurances to survey participants that substantially limit the ability to share the microdata with other programs within the same or other agencies. In still other cases, laws governing data collected and controlled by federal agencies restrict the access and use of the data.[22,23] In general, the authorized uses include only public health research, and/or the stated purposes for which the data were collected, and only with the explicit consent of the individual (or establishment) from which it was collected. As one person stated, “The terms of sharing data are in the data use agreement. It requires that the requestor tell us where they plan to store the data, who will be responsible for the data, and that the data be destroyed upon completion of the use”. Other agency representatives shared the same information. “To use the data for anything other than the extensive tabular data provided by the [agency supplying data] one has to go through a process that meets the requirements for all of those regulations. This also includes not only the analysis being of benefit to you, but it has to be of benefit to the [sending or source agencies]. It is a rigorous process to get access to the data other than the one we get yearly.”

[22] Government Publishing Office. (n.d.). Title 42 -The Public Health and Welfare. Retrieved from https://www.gpo.gov/fdsys/pkg/USCODE-2014-title42/pdf/USCODE-2014-title42-chap6A-subchapII-partA-sec242m.pdf

[23] U.S. Government Printing Office. (1996, August 21). Health Insurance Portability and Accountability Act of 1996, Public Law 104-191. Retrieved from https://www.gpo.gov/fdsys/pkg/PLAW-104publ191/html/PLAW-104publ191.htm

Challenge 5
Norms & Resource Constraints

Data representatives do not see the demand for sharing restricted and nonpublic data; view the public use files as sufficient for the majority of analyses; and, for certain data programs, view data sharing requests as ad-hoc or special. Strained resources, fear of misrepresentation of the data, and reluctance to critique a sister agency for unsatisfactory data sharing practices all contribute to maintaining the status quo.

Agencies want to, but are unable to, quickly discover restricted and nonpublic data outside their agency

Information about data assets and sources for internal employees is limited. Among the data representatives and leaders who were interviewed, most expressed a lack of awareness yet were curious about what data HHS collects and what data may be beneficial for an analytical project or programmatic activity. This lack of knowledge has contributed to a perceived lack of demand for restricted and nonpublic data sharing.

In interviews, several data representatives expressed, “I did not know what data exists [across agencies]”. This presents challenges for agencies who want to explore what data might be relevant to their work. It also causes inefficiencies as analysts spend considerable time locating data sources relevant to their project and identifying the person who governs access to the data. Some analysts suggested there needs to be greater development, dissemination, and use of the Enterprise Data Inventory for internal and external audiences.

Depending on the type of data they work with, staff may not view data sharing as part of their job

Willingness and concern about increasing data sharing among agencies vary depending on whether the data are created by and for statistical purposes or created as a byproduct of administrative activities. For statistical datasets, agency personnel consistently expressed a commitment to making data publicly available in ways that respect privacy and confidentiality. One agency staff member said, “Data is like manure. It is best when it is spread around.” Representatives of statistical data assets expressed the belief that the public use files should be the starting point for most analyses. These public use files are available to anyone.

For administrative data, sharing data is not the agency’s primary purpose. A data analyst described her experience requesting administrative data, “There is a feeling that [sharing nonpublic data] is not part of [the data representative’s] job. [Sharing data] was viewed as a special request and was treated as such.” These sentiments stem from resource constraints, as well as a risk-averse environment, given legal requirements concerning privacy and unintended disclosures. Certainly, these sentiments play a role in shaping and reinforcing a culture limiting data sharing.

Agency resources to implement data sharing are thin

While public use files have been prioritized as an output from data, resources available to make restricted or nonpublic data available vary greatly across the agencies; across the board, agencies are at or below capacity. One agency personnel remarked, “We cannot advertise the availability of the data to other agencies because we don’t have enough

expertise [or resources] to process the requests when they come in". At least 7 agencies reported their staffing levels for data collection efforts as not sufficient to meet increased demand for data sharing. In speaking about low staffing levels, one agency clarified "We don't [just] need staff; we need expertise," referring to the expertise to deeply understand the data, its lifecycle, and the scope of its capacity for analyses.

Agencies are careful not to burden or critique other agencies for how they treat data requests

As stated previously, even when a requesting agency needs data and knows of a source within the department, they may not "want to burden sister agencies" by requesting data. Agencies supplying the data can be reluctant to appear to favor one request over another. Agencies also do not want to take resources away by inundating sister agencies with data requests. Limited resources and funds make agencies hesitant to ask others for increased data collaboration.

When asked about how the system can be improved, agency personnel clearly understand the resources and demands placed on their sister agencies and are hesitant to critique them. Reflecting on the costs of acquiring data from a different agency, one agency remarked, "We understand how much money goes into data collection and management," with hesitation to express frustration with the process or costs of acquiring data. During interviews, analysts frequently withheld the names of which data were difficult to request and acquire.

NEXT STEPS

The U.S. Department of Health and Human Services, at both the leadership and staff levels, has dedicated an enormous amount of time discussing the current state of data sharing across the organization. While perhaps not an exhaustive approach, the process has been representative of the range of challenges and potential opportunities to enhance how data is shared across HHS agencies. Understanding this landscape is only the first step.

As the Department begins to address these challenges, identifying use cases and demonstrating the business value of data sharing will be critical. This will provide grounds for harmonizing data governance into a central function as one of the first priorities. Creating a robust technical environment for data analysis, workflow management, and streamlining data acquisition will be essential. Attention should be given to existing interagency and data use agreements that relate to data collection and use as well as relevant regulatory reform. While this is a long-term effort, an evaluation of next steps has begun. Underpinning each of these changes is the need for building workforce capacity. Already, data science training programs have been initiated, and demand from staff has been very positive.

Data sharing is a fundamental instrument of collaboration, which can lead to a more effective and efficient organization. Each of the areas highlighted in this report will need to be incrementally but persistently addressed. If data is to be leveraged as an asset using advanced analytic tools and predictive modeling, the use of data must be essential to a Departmental strategy rather than purely individual project based. Efforts are underway to construct an enterprise-wide data sharing framework, through validation and collaboration with agencies and using an agile development approach. Ultimately, success will require a long-term investment, continued collaboration, and the iterative demonstration of value from data to drive the culture change essential to transforming HHS.

APPENDIX

List of Agencies & Data Assets

HHS Agency	Data Asset Name	Functional Description	Website
Administration for Children and Families	Adoption and Foster Care Analysis and Reporting System	The Adoption and Foster Care Analysis and Reporting System (AFCARS) collects case-level information from state and tribal title IV-E agencies on all children in foster care and those who have been adopted with title IV-E agency involvement. Examples of data reported in AFCARS include demographic information on the foster child as well as the foster and adoptive parents, the number of removal episodes a child has experienced, the number of placements in the current removal episode, and the current placement setting.	https://www.acf.hhs.gov/cb/research-data-technology/reporting-systems/afcars
Administration for Children and Families	National Directory of New Hires	The federal Office of Child Support Enforcement (OCSE) operates the National Directory of New Hires (NDNH), a database established by the Personal Responsibility and Work Opportunity Reconciliation Act of 1996 (PRWORA) for the purposes of assisting state child support agencies in locating parents and enforcing child support orders. In addition, Congress authorized specific state and federal agencies to receive information from the NDNH for authorized purposes.	https://www.acf.hhs.gov/css/resource/a-guide-to-the-national-directory-of-new-hires
Administration for Community Living	National Adult Maltreatment Reporting System	The National Adult Maltreatment Reporting System (NAMRS) is the first comprehensive, national reporting system for adult protective services (APS) programs. It collects quantitative and qualitative data on APS practices and policies, and the outcomes of investigations into the maltreatment of older adults and adults with disabilities.	https://www.acl.gov/programs/elder-justice/national-adult-maltreatment-reporting-system-namrs

HHS Agency	Data Asset Name	Functional Description	Website
Administration for Community Living	State Health Insurance Assistance Program	The State Health Insurance Assistance Program (SHIP) provides Medicare beneficiaries with information, counseling, and enrollment assistance. Its mission is to strengthen the capability of grantees to support a community-based, grassroots network of local SHIP offices that assist beneficiaries with their Medicare-related questions.	https://www.acl.gov/programs/connecting-people-services/state-health-insurance-assistance-program-ship
Agency for Healthcare Research and Quality	Healthcare Cost and Utilization Project	The Healthcare Cost and Utilization Project (HCUP) is the Nation's most comprehensive source of hospital care data, including information on inpatient stays, ambulatory surgery and services visits, and emergency department encounters. HCUP enables researchers, insurers, policymakers and others to study health care delivery and patient outcomes over time, and at the national, regional, State, and community levels.	https://www.ahrq.gov/research/data/hcup/index.html
Agency for Healthcare Research and Quality	Medical Expenditure Panel Survey	The Medical Expenditure Panel Survey (MEPS) is a set of large-scale surveys of families and individuals, their medical providers, and employers across the United States.	https://meps.ahrq.gov/mepsweb/
Centers for Disease Control and Prevention	Behavioral Risk Factor Surveillance System	The Behavioral Risk Factor Surveillance System (BRFSS) is the nation's premier system of health-related telephone surveys that collect state data about U.S. residents regarding their health-related risk behaviors, chronic health conditions, and use of preventive services. Established in 1984 with 15 states, BRFSS now collects data in all 50 states as well as the District of Columbia and three U.S. territories. BRFSS completes more than 400,000 adult interviews each year, making it the largest continuously conducted health survey system in the world.	https://www.cdc.gov/brfss/about/index.htm
Centers for Disease Control and Prevention	National Notifiable Diseases Surveillance System	The National Notifiable Diseases Surveillance System (NNDSS) is a nationwide collaboration that enables all levels of public health—local, state, territorial, federal, and international—to share notifiable disease related health information.	https://wwwn.cdc.gov/nndss/

HHS Agency	Data Asset Name	Functional Description	Website
Centers for Disease Control and Prevention	National Syndromic Surveillance Program	The National Syndromic Surveillance Program (NSSP) promotes and advances development of a syndromic surveillance system for the timely exchange of syndromic data. These data are used to improve nationwide situational awareness and enhance responsiveness to hazardous events and disease outbreaks to protect America's health, safety, and security. NSSP functions through collaboration among individuals and organizations at local, state, and federal levels of public health; federal agencies including the U.S. Department of Defense and the U.S. Department of Veterans Affairs; public health partner organizations; and hospitals and health professionals.	https://www.cdc.gov/nssp/
Centers for Disease Control and Prevention, National Center for Health Statistics	National Hospital Ambulatory Medical Care Survey	The National Ambulatory Medical Care Survey (NAMCS) is designed to meet the need for objective, reliable information about the provision and use of ambulatory medical care services in the United States. Findings are based on a sample of visits to nonfederal employed office-based physicians who are primarily engaged in direct patient care and, starting in 2006, a separate sample of visits to community health centers.	https://www.cdc.gov/nchs/ahcd/index.htm
Centers for Disease Control and Prevention, National Center for Health Statistics	National Health and Nutrition Examination Survey	The National Health and Nutrition Examination Survey (NHANES) is a program of studies designed to assess the health and nutritional status of adults and children in the United States. The survey is unique in that it combines interviews and physical examinations.	https://www.cdc.gov/nchs/nhanes/index.htm
Centers for Disease Control and Prevention, National Center for Health Statistics	National Health Interview Survey	The National Health Interview Survey (NHIS) has monitored the health of the nation since 1957. NHIS data on a broad range of health topics are collected through personal household interviews. For over 50 years, the U.S. Census Bureau has been the data collection agent for the National Health Interview Survey.	https://www.cdc.gov/nchs/nhis/index.htm

HHS Agency	Data Asset Name	Functional Description	Website
Centers for Disease Control and Prevention, National Center for Health Statistics	National Vital Statistics System	The National Vital Statistics System contains National Vital Statistics, tracking nationwide "vital events" defined as births, deaths, marriages and divorces.	https://www.cdc.gov/nchs/nvss/index.htm
Centers for Medicare and Medicaid Services	Medicare Fee-for-Service Claims	Medicare Fee-for-Service (FFS) claims data include information submitted to CMS by healthcare providers for payment for services provided to beneficiaries enrolled in Medicare Parts A and/or B. These data include dates of service, diagnosis/procedure codes, and payment information, among other variables.	https://www.resdac.org/cms-data?tid%5B%5D=4931
Centers for Medicare and Medicaid Services	Medicare Advantage Encounter	Medicare Advantage Encounter data are detailed data submitted to CMS by health plans with information on healthcare services provided to beneficiaries enrolled in the Medicare Advantage program. These data include dates of service and diagnosis/procedure codes, among other variables.	https://www.resdac.org/cms-data?tid%5B%5D=6056
Centers for Medicare and Medicaid Services	Part D Prescription Drug Events	Medicare Part D Prescription Drug Event data are submitted to CMS by PDP sponsors and include information on prescription drug fills. These data include the drug code (NDC), fill date, and gross drug costs, among other variables.	https://www.resdac.org/cms-data?tid%5B%5D=6066
Food and Drug Administration	Adverse Event Reporting System	The FDA Adverse Event Reporting System (FAERS) is a database that contains adverse event reports, medication error reports and product quality complaints resulting in adverse events that were submitted to FDA. The database is designed to support the FDA's post-marketing safety surveillance program for drug and therapeutic biologic products.	https://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/

HHS Agency	Data Asset Name	Functional Description	Website
Food and Drug Administration	Sentinel	The U.S. Food and Drug Administration's (FDA) Sentinel Initiative is a long-term effort to improve the FDA's ability to identify and assess medical product safety issues. The Sentinel System is an active surveillance system that uses routine querying tools and pre-existing electronic healthcare data from multiple sources to monitor the safety of regulated medical products. FDA-Catalyst activities leverage the Sentinel Infrastructure by utilizing the data available through its Data Partners and supplementing it with data from interventions or interactions with members and/or providers.	https://www.sentinelinitiative.org/
Health Resources and Services Administration	BHW Management Information System Solution	The BHW Management Information System Solution is a multi-module system comprised of three separate modules that together support the mission of HRSA's Bureau of Health Workforce (BHW). The BHW's ultimate goal and key outcome is to enable access to medical services for over 12.2M patients from underserved and rural communities.	https://www.hrsa.gov/about/organization/bureaus/bhw/index.html
Health Resources and Services Administration	Health Center Patient Survey	The Health Center Patient Survey (HCPS), sponsored by HRSA, provides robust patient-level data to determine how well health centers funded under Section 330 of the Public Health Service Act provide access to primary and preventive health care.	https://bphc.hrsa.gov/datareporting/research/hcpsurvey/index.html
Health Resources and Services Administration	National Practitioner Data Bank	This web-based repository of reports is used as a workforce tool to enhance professional review efforts, and prevent health care fraud and abuse, with the ultimate goal of protecting the public.	https://www.npdb.hrsa.gov/
Health Resources and Services Administration	Uniform Data System	Each year HRSA-funded health center grantees are required to report a core set of information, including data on patient demographics, services provided, clinical indicators, utilization rates, costs, and revenues.	https://bphc.hrsa.gov/datareporting/reporting/index.html
National Institutes of Health	Unfunded Research Grants	eRA provides critical IT infrastructure to manage over $30 billion in research and non-research grants awarded annually by NIH and other grantor agencies in support of the collective mission of improving human health. eRA systems, including eRA Commons, ASSIST and IMPAC II modules, support the full grants life cycle and are used by applicants and grantees	https://era.nih.gov/

HHS Agency	Data Asset Name	Functional Description	Website
		worldwide as well as federal staff at the NIH, AHRQ, the CDC, FDA, SAMHSA, and VA	
Substance Abuse and Mental Health Services Administration, Center for Behavioral Health Statistics and Quality	Mental Health Client-Level Data	The Mental Health Client-Level Data is mental health client-level data that comes from the states, District of Columbia, and US Territories. Client-level data includes a limited set of demographic, clinical attributes, and outcomes routinely collected in monitoring individuals receiving mental health and support services.	https://wwwdasis.samhsa.gov/dasis2/mhcld.htm
Substance Abuse and Mental Health Services Administration, Center for Behavioral Health Statistics and Quality	Treatment Episode Data Set	The Treatment Episode Data Set (TEDS) comprises of data that is routinely collected by States in monitoring their individual substance abuse treatment systems. In general, facilities reporting TEDS data are those that receive State alcohol and/or drug agency funds (including Federal Block Grant funds) for the provision of substance abuse treatment.	https://wwwdasis.samhsa.gov/webt/information.htm
Substance Abuse and Mental Health Services Administration	National Survey on Drug Use and Health	The National Survey on Drug Use and Health provides national and state-level data on the use of tobacco, alcohol, illicit drugs (including non-medical use of prescription drugs) and mental health in the United States.	https://www.samhsa.gov/data/data-we-collect/nsduh-national-survey-drug-use-and-health

List of Acronyms

Acronym	Definition
CDC	Centers for Disease Control and Prevention
CFR	Code of Federal Regulations
CIPSEA	Confidential Information Protection and Statistical Efficiency Act
CMS	Centers for Medicare and Medicaid Services
DSA	Data Sharing Agreement
DUA	Data Use Agreement
EPPE	Enterprise Privacy Protection Engine
eRA	Electronic Research Administration
FAERS	FDA Adverse Event Reporting System
FDA	U.S. Food and Drug Administration
FFS	Medicare Fee-for-Service Claims
HCPS	Health Center Patient Survey
HCUP	Healthcare Cost and Utilization Project
HHS	U.S. Department of Health and Human Services
HIPAA	Health Insurance Portability and Accountability Act
HRSA	Health Resources and Services Administration
IAA	Interagency Agreement
ICD	International Classification of Diseases
MEPS	Medical Expenditure Panel Survey
MOU	Memorandum of Understanding
NAMCS	National Ambulatory Medical Care Survey
NAMRS	National Adult Maltreatment Reporting System
NCHS	National Center for Health Statistics
NDC	National Drug Code
NDNH	National Directory of New Hires
NHANES	National Health and Nutrition Examination Survey

Acronym	Definition
NHIS	National Health Interview Survey
NIH	National Institutes of Health
NNDSS	National Notifiable Diseases Surveillance System
NSSP	National Syndromic Surveillance Program
NYTD	National Youth in Transition Database
OCSE	Office of Child Support Enforcement
PII	Personality Identifiable Information
PMA	President's Management Agenda
PRWORA	Personal Responsibility and Work Opportunity Reconciliation Act
RDC	Restricted Data Center
SAMHSA	Substance Abuse and Mental Health Services Administration
SHIP	State Health Insurance Assistance Program
TEDS	Treatment Episode Data Set