2018 HHS Data Strategy: Enhancing the HHS Evidence-Based Portfolio

A Report from the U.S. Department of Health and Human Services Data Council

The U.S. Department of Health and Human Services Data Council

CONTENTS

FIGURES

TABLES

EXECUTIVE SUMMARY

The Department of Health and Human Services (HHS) is one of the largest federal agencies, with 11 Operating Divisions, 16 Staffing Divisions, and 10 Regional Offices. HHS data provide the capacity to monitor and improve the nation's health care system, human services programs, and population health. The Department's data collection systems include the full spectrum of information on health and health care, ranging from demographic and public health trends to the quality of care for individuals of all ages across the United States. To achieve its mission, it is crucial for HHS to collect, analyze, and disseminate high-quality, reliable data that can inform policymaking.

As the principal internal advisory body to the Secretary of Health and Human Services on the Department's data and statistical policy, the HHS Data Council develops, implements, and updates the Department's data strategy. In doing so, the Council periodically assesses the Department's survey and data collection portfolio to identify data collection strategies in order to expand the capacity of HHS's data resources; promote synergy across data systems; ensure the efficiency, quality, utility, and timeliness of data collection systems; and address high-priority gaps in data.

The availability, variety, and attention garnered by data have increased dramatically since the development of the Department's last data strategy in 2011. The President's Management Agenda,[1] Reimagine HHS,[2] the HHS Agency Priority Goal: Combined Data Analysis,[3] and the Report of the Commission on Evidenced-Based Policymaking (CEP) have all called for leveraging data to provide insight into the effectiveness of programs and to inform decision making. Given the vision provided by these initiatives and in recent reports, this is an ideal time to examine the HHS data strategy and the role of the Data Council in opening up opportunities for making the best use of HHS data. Building on the previous departmental data strategy and on other internal and external reports, the 2018 HHS Data Strategy presented in this report focuses on enhancing the Department's evidence-based portfolio to improve the use of data for policy research and program evaluation. As outlined below, the 2018 HHS Data Strategy has six priority areas in which there are opportunities to pursue as well as strategies that can advance these areas.

PRIORITY 1: IMPROVING ACCESS TO HHS DATA

Opportunity 1: Increase the accessibility of HHS data to internal and external users while ensuring that the information is used responsibly

Strategy 1: Streamline processes for accessing data

Opportunity 2: Increase awareness within the Department about available HHS data resources and research

Strategy 1: Develop and implement a framework for a catalog of HHS Data resources

Strategy 2: Establish a process to coordinate the dissemination of major new data releases and research briefs across HHS

PRIORITY 2: ENHANCING ADMINISTRATIVE DATA FOR RESEARCH

Opportunity 1: Expand the use of administrative data in the Department

Strategy 1: Improve the documentation and curation of HHS administrative data

Opportunity 2: Improve the quality of administrative data for research

Strategy 1: Develop quality frameworks for administrative data collection

Strategy 2: Create procedures to benchmark big data for program evaluation

PRIORITY 3: INCREASING DATA LINKAGES ACROSS DIVERSE DATA ASSETS

Opportunity 1: Apply existing departmental knowledge and lessons learned from data linkages

Strategy 1: Develop an HHS Data Linkage Repository that includes information on linkage strategies, barriers, and opportunities

Opportunity 2: Improve the capacity to link HHS data internally and with other data sources

Strategy 1: Promote data linkage between HHS agencies and between HHS and other federal agencies to address Departmental priorities

Strategy 2: Promote data linkage to nonfederal data

PRIORITY 4: MODERNIZING PRIVACY PROTECTIONS

Opportunity 1: Increase data sharing without eroding privacy protections through better communication and coordination with experts

Strategy 1: Use data intermediaries to facilitate data sharing in accordance with privacy laws

Strategy 2: Increase the use of disclosure review boards or data disclosure boards and provide guidance on best practices for de-identification

Opportunity 2: Assist in standardization of departmental privacy policy practices

Strategy 1: Support the Department's development of more streamlined data sharing processes, including model enterprise-wide DUAs and inter-agency agreements (IAA)

Strategy 2: Address privacy and legal concerns about the use of data for policy research, statistical purposes, and program evaluation

PRIORITY 5: INCREASING DATA POLICY COORDINATION AND INFORMATION SHARING ACROSS THE DEPARTMENT

Opportunity 1: Increase coordination in the Department regarding data collection, system and software investments, and data management and governance

Strategy 1: Identify high-priority data and information policy issues that require OS-level coordination

Strategy 2: Increase communication between the Data Council, OCIO, ONC, and the CTO

Opportunity 2: Inform policymakers and researchers about the value and uses of HHS data

Strategy 1: Communicate the value of HHS data collections and systems for policymaking

Strategy 2: Inform internal stakeholders about potential tradeoffs between timeliness and quality and how this impacts the usefulness of data for policy purposes

PRIORITY 6:   BUILDING A 21ST CENTURY DATA-ORIENTED WORKFORCE

Opportunity 1: Enhance the data science capability of the current HHS data workforce

Strategy 1: Increase data science and statistical training opportunities

Strategy 2: Promote multidisciplinary data science teams and increase cross-program collaboration

Strategy 3: Promote awareness and education of data ethics in the Department

Opportunity 2: Reinforce capacity to explore the application of data science and alternative data to HHS research and program evaluation

Strategy 1: Develop capacity to investigate new or more-blended statistics for health and human services

Strategy 2: Develop the capacity to coordinate the evaluation of alternative data sources

Strategy 3: Explore the NIH Strategic Plan for Data Science as a tool to support data science across the Department

Opportunity 3: Invest in the future of data science

Strategy 1: Increase the number of new data scientists

Strategy 2: Ensure that staff have the expertise to explore the coordinated implementation of technology or software that facilitates ethical data sharing and use for data science capabilities

By enhancing the wealth of health data collected through the its programs, HHS will be better equipped to address the nation's most pressing policy and public health challenges—including stemming the opioid epidemic; rigorously evaluating and supporting innovation in the programs and services that the Department provides; working effectively across sectors to address human services needs; and identifying ways to reduce the cost of health insurance and prescription drugs while improving the quality of care, to name a few.

INTRODUCTION – ENHANCING THE EVIDENCE-BASED PORTFOLIO

The 2018 HHS Data Strategy focuses on improving the Department's capacity to develop statistical evidence to support policymaking and program evaluation. For the purposes of this data strategy, evidence is defined as "information produced by 'statistical activities' with a 'statistical purpose' that is potentially useful when evaluating government programs and policies."[4] Both the Office of Management and Budget (OMB) and advisory committees have asked federal agencies to do the following: increase the use of their data for policy research and program monitoring and evaluation, improve the availability and timeliness of data for decision making, and augment the usefulness of federal data by linking the information to other sources of data to fill knowledge gaps. HHS is already at the forefront of using data to build the evidence base that supports health care policy research and program evaluation. The CEP recognized the HHS Data Council as a leader in successfully coordinating data resources within a federal agency.

The HHS Data Council was established in 1996 to better coordinate data and statistical policy in the Department across Operating Divisions (OpDivs) and Staffing Divisions (StaffDivs). The leadership of the Council is responsible for developing a long-term departmental data strategy to address current challenges related to using data for evidence-building. These challenges include, but are not limited to, improving data curation and archiving, augmenting the use of administrative data and data linkages, modernizing privacy protections through increased investment in technological solutions for secure access to data, improving policy coordination to better align HHS systems for data sharing between OpDivs, and investing in a workforce that can navigate the data science environment. The value of HHS data depends on its relevance, timeliness, availability, and distinctive contributions. The Data Council has developed guidelines to preserve and advance the value of HHS data for policymaking (a fuller discussion of this effort appears in Appendix A.)

The 2018 HHS Data Strategy is aligned with recommendations from the CEP; the National Academies of Sciences, Engineering, and Medicine (NAS); and federal initiatives that are intended to improve the use of data for policy research and program evaluation. The CEP was a congressionally mandated committee charged with identifying how the federal government can use its existing data to improve programs and policies. In September 2017, the CEP published its report,[5] which included recommendations for strengthening federal evidence-building capacity while improving access to secure, private, and confidential data. Additionally, the NAS released a series of panel reports that provided recommendations on improving the utility of federal data for evidence-based policymaking.

The Executive Branch has recognized that the wealth of government data is a strategic asset. The President's Management Agenda,[6] under the Cross-Agency Priority (CAP) Goal "Leveraging Data as a Strategic Asset," has highlighted the importance of using data to inform decision making, policymaking, and oversight, and to foster innovation and learning. As part of that effort OMB is developing a federal data strategy that will focus on four areas: (1) enterprise data governance; (2) access, use, and augmentation; (3) decision making and accountability; and

(4) commercialization, innovation, and public use. HHS has worked to support the President's Management Agenda through its own Agency Priority Goal, which is to increase the combined analysis of disparate data sets in order to develop deeper insight into evidence-based policymaking. This goal is aligned with the ReImagine HHS strategic shift area.[7] Additionally, OMB Memorandum M-14-06 has directed federal agencies to increase the use of administrative data for statistical purposes and to report on progress as part of the evidence they submit during the annual budget process.[8]

Many of the principles and recommendations outlined in the reports cited above have already been implemented by HHS. The data generated by program offices and statistical agencies across HHS provide evidence that is valuable in terms of forming policies that support the Department's priorities.[9] Accurate, timely, comparable data provide an objective basis for determining the effectiveness of programs and services, and the allocation of resources and future investments. Assessing such evidence provides an opportunity to improve programs that are not meeting expectations or to understand the best practices of programs that show significant results. The 2018 HHS Data Strategy will support data needs in high-priority health areas identified in the 2018-2022 HHS Strategic Plan.[10] That plan highlights the need to enhance real-time and local data collection, and to implement information technology (IT) solutions and innovative methods that support the timely exchange of information between HHS and nonfederal agencies.

Framed around increasing the Department's evidence-based portfolio, the 2018 HHS Data Strategy has three objectives. First, it will address data collected by or for various HHS agencies, including not only survey data but also various forms of administrative data that are potentially useful for monitoring population health, evaluating government programs, and informing policies that address the health of the nation. Second, the strategy is intended to build upon efforts to further the use of data for evidence-building. Third, the strategy is designed to complement data activities across the Department and the government, and to demonstrate the alignment of such activities. Focusing on generating evidence will help the Department to achieve its mission-critical goals of addressing emerging public health challenges and improving the quality of the nation's health and human services programs.

PRIORITY 1 – IMPROVING ACCESS TO HHS DATA

A. Background

Improving access to HHS data while maintaining its quality and confidentiality is key to expanding the Department's ability to generate evidence that can inform policy and programs. This evidence may come from a variety of programs. Access to the programs' data and combining data from multiple sources form the basis for developing effective, evidence-based policy and programs. This section covers the types of access that are granted to users and the methods through which they can access data.

Types of data access

Access to data—particularly to administrative data, which may not be designed for research purposes or distribution—encompasses a range of issues, including determining availability,

receiving the data, merging multiple data sets, and understanding what the data mean.[11] Restrictions on access can involve modifying and masking data sets and systems to maintain confidentiality and minimize the risk of disclosure. This section discusses physical access to data, both for external stakeholders (the public) and internal stakeholders (government personnel and contractors).

Public access

HHS agencies make their data available to the public in a variety of ways. Tools for accessing data, such as CDC Wonder[12] and HCUPnet,[13] allow users to tabulate data without having the ability or resources to analyze individual-level data. Some HHS agencies produce de-identified, microdata public use files (PUFs) for the general public to download. These files are generally available at no charge on government websites or on individual agency sites. HHS agencies also provide metadata on the location and availability of data sets across the Department. In addition, agencies often provide access to public use data through advanced, online data query tools, visualization platforms, and data dashboards which may allow for the selection of data points, tabulations, and calculation of other statistical inquiries. Finally, agencies provide access to their research; to their staff's research; and to annual reports, briefs, special topics reports, and numerous graphs, tables, and data visualizations on their websites.

Restricted access

Although PUFs are often freely available on government websites, restricted use files (RUFs) may be accessed through a variety of means. These files are also far more detailed than PUFs because they may contain personally identifiable information (PII)[14] or protected health information (PHI).[15] Depending on the requirements of the specific data system, access to RUFs can be limited to secure research environments or is granted through vehicles such as virtual data enclaves. These enclaves are virtual machines launched from users' desktops providing secure remote access to environments that are sometimes housed in data archiving entities, authorized universities, or contracts developed directly between organizations and HHS.[16] To access these files for secondary analysis, applicants will often have to prove that they are affiliated with a research institution and provide a project proposal that describes the project for which they need the data. NCHS developed Research Data Centers (RDCs)[17] to allow researchers access to restricted data in the most secure type of controlled environment. In addition to providing access to NCHS data, the RDCs also host restricted data from a variety of groups within HHS. The RDC system is responsible for protecting the confidentiality of survey respondents, study subjects, and institutions from which data were collected. An RDC does not transmit data sets electronically but gives users options for accessing data on site and remotely.[18] In recent years, RDCs have expanded to numerous locations.

HHS: Internal and interagency access

There are different ways to access public and restricted use data within HHS. The process for accessing public use data usually entails downloading PUFs from the agency websites or using online data tools to analyze public use data. However, there is variability across OpDivs in the processes for accessing restricted use data. Furthermore, at the Department level, there is no

standard repository from which users can access research results; nor are there shared standards for data management and archiving, or standard requirements or enforcement of practices for agency-to-agency sharing of data.[19] HHS agencies do maintain individual interagency agreements (IAAs), memorandums of understanding (MOUs), and data use agreements (DUAs) with sister agencies within the Department, but these arrangements vary, and they are distinct from each other.

The Department is working to improve its internal processes for accessing data across its agencies. This would allow for the timely sharing of information that is needed to inform urgent public health issues. Streamlining and standardizing these processes would also help programs use departmental data to inform decision making. Increasing access to HHS data for Department staff would also raise their awareness of the data sets that are available across the Department, leading to more collaborative efforts and integrated research findings.

B. Issues

Data the Department collects can be effectively leveraged to yield deeper insight into issues than originally intended, particularly when data from multiple sources are combined. Although the Department already makes great use of the administrative data it collects in the specific programs for which it was collected, sharing data across programs is impeded by several issues, both real and perceived. Similar concerns arise in sharing RUFs or any data sets that have not been released as PUFs.

Legal restrictions

Specific data sets and some general data elements are prohibited by law from being shared, or they can only be shared when certain conditions are met. However, these laws are not always well understood by the average data curator, leading many to take the safer route of not sharing data. Another restriction may be the conditions of the informed consent given by study participants, which could dictate who can use the data and whether it can be linked to other data.

Perceptions of data access processes

When access to data is restricted but not prohibited by law, the processes for controlling access may not be well understood, or they may be loosened following a review. Most policies that govern access to data are managed at the OpDiv level.

Lack of widespread knowledge of departmental data resources

The data sets available to HHS analysts are not widely known across the Department. There is no reliable, comprehensive, centralized inventory of data sets that analysts can reference to obtain information on all departmental data sets available for research purposes. Moreover, previous and current efforts to provide such a resource have always been well-subscribed.[20]

Difficulty of linking to other data sets

Researchers and policymakers may learn of data from several HHS programs or agencies that have a bearing on their topics of interest. However, often the structure of available data sets is

not clearly described. As a result, even when data sets from multiple sources are discovered, it is often difficult to find an element for linking them. The quality and consistency of identifiers across data sets can be a major challenge in linking data.

Cost

It can be difficult to provide data because of the cost associated with collecting, curating, and maintaining desired levels of accessibility to the information. HHS and non-HHS analysts alike may be required to pay a user fee to the providing agency to cover the costs of making the data available for some HHS data. Furthermore, it will be difficult for HHS to continually make data accessible to researchers in the absence of appropriate funding mechanisms that (1) ensure that vital data collection activities coincide with the push toward evidence-based policymaking and then (2) allow the data to be made available.

C. Opportunities

The following two opportunities are proposed to improve access to HHS data in support of enhancing the Department's evidence-based portfolio.

Opportunity 1: Increase the accessibility of HHS data to internal and external users while ensuring that the information is used responsibly

Strategy 1: Streamline processes for accessing data

A current obstacle to data access for research purposes is the lack of consistent policies and processes between HHS agencies. Streamlining the processes for accessing data means aligning and coordinating existing and future data sets. HHS agencies can work to merge current and future data sources by unifying standard data access policies and data collection processes. Additionally, HHS agencies can be responsible for sufficiently documenting and managing data through its lifecycle, leading to uniformity in and the expeditious release of the data. The HHS 2018-2022 Strategic Plan outlines a digital strategy to modernize access to data, including cloud computing and database consolidation.[21]

Opportunity 2: Increase awareness within the Department about available HHS data resources and research

Strategy 1: Develop and implement a framework for a catalog of HHS data resources

HHS could develop and implement a pilot departmental data catalog of PUFs and RUFs that would be accessible through a proposal process. The catalog will help to create more knowledge of HHS data in the Department and a sustainable mechanism for updating that information. The catalog would be designed to give users key information about the quality of the data and the contents of the data files (possibly including information on the linkage potential of the files). In addition, the data catalog would provide some means of collaboratively documenting these metadata about the data set, allowing it to grow over time.

It must be recognized that IT and other resources will be required to administer a data catalog, but those costs may be minimized through creative collaboration within existing infrastructures

and processes. This may include following standards set out by OMB guidance in Memorandum M-13-13 on Open Data[22] and tapping into work within the Department that is meant to automate current reporting requirements, working to build the catalog into existing collaborative sites like the Department's strategic planning system, and following methods meant to minimize or eliminate the need for dedicated HHS staff through crowdsourcing methodology. The Data Council could lead a key first step to determine which data files should be included in the catalog and the information about the files that should be required. The Council could also advise on departmental efforts to redesign of the HHS Enterprise Data Inventory (EDI).[23] This could (1) improve the search capability and the metadata that document administrative data collections and (2) make data more accessible and user-friendly for research.

Strategy 2: Establish a process to coordinate the dissemination of major new data releases and research briefs across HHS

While HHS is constantly releasing important data useful to a wide range of interests, some releases rise to the level of informing major policy decisions or providing new information affecting the American population. HHS could promote the cross-agency release of new data to increase awareness of HHS data resources in these high-visibility priority areas. To serve the needs of external users, the Department could create a more centralized and accessible platform where users could access newly released public data. Given the volume of data and reports generated by HHS, initial coordination efforts will likely be limited to high-priority areas.

PRIORITY 2 – ENHANCING ADMINISTRATIVE DATA FOR RESEARCH

A. Background

Administrative data consist of "information collected from individual persons, businesses, or institutions for the purpose of managing programs, implementing regulations, enforcing laws, or other purposes that affect these entities."[24,25] Although administrative data are primarily used for program monitoring and similar purposes, they are increasingly serving as sources of evidence-based policymaking for federal programs. HHS collects administrative data for several primary purposes, including program and grant monitoring and management, policymaking, enforcement in areas such as child support and program fraud, and surveillance. As a secondary purpose, administrative data are used for statistical research. In some cases, HHS collects administrative data from states for a statistical purpose. Administrative data must be developed or curated, with the appropriate documentation quality checks, and appropriate privacy protections, to be used for statistical purposes.

To ensure that administrative data are used appropriately for statistical purposes, they must be legally protected. This requires an additional investment in staff training and in the development of internal policies that govern the use of the data. Some administrative data collections are also governed by statutes limiting their use.[26] These statutes enable HHS and other government entities to collect sensitive, valuable data while protecting the privacy and confidentiality of individuals and organizations, and safeguarding their data against potential misuse.[27]

The value of administrative data

There has been a recent push to use administrative data for evidence-based policymaking not only because of the value of the information, but also for making the best use of federal resources. The OMB memorandum, *Guidance for Providing and Using Administrative Data for Statistical Purposes* (M-14-06), states:

> Managing administrative data with statistical purposes in mind will reduce burden on the public by making use of information about individuals, businesses, and institutions that would otherwise need to be collected through surveys or would simply be unobtainable from surveys within an acceptable level of burden or accuracy.[28]

HHS has also invested in enhancing the use of administrative data for research. The Data Council created the HHS Administrative Data Working Group (ADWG**)** to identify better ways of using HHS administrative data collections.

Although administrative data can be quite variable along dimensions such as quality, timeliness, and content, they have many advantages that make them useful for research and program evaluation. For instance, when administrative records include all occurring events, there is little collection bias, and the data provide analysts with the opportunity to make more accurate estimates of relatively rare events. Thus, the secondary use of routinely collected administrative data can be cost-effective compared with primary data collection. At the same time, the limitations of administrative data must be recognized, including the possibility that the universe for the administrative program may not include the entire universe of interest, and the information can be less detailed, especially information that was not necessary to the original purpose of the data. Although the careful choice of analytic tools is always important, the sheer size of some administrative data sets can lead to misinterpreting statistically significant results. Administrative data can also can assist survey developers in constructing sampling frames and in providing auxiliary information for editing, imputation, and estimation.

In recent years, there have been calls for the federal government to make greater use of its administrative data collections. In response, the Data Council formed the ADWG and produced an environmental scan of HHS administrative data sets.[29] The report first identified the primary purposes for which agencies collect administrative data. Second, it analyzed the 29 "high-value" administrative data sets across agencies that were selected by ADWG. Third, it suggested challenges and next steps for solutions. The next steps included (1) cataloging HHS data inventory and building useful metadata, (2) improving data quality, (3) better managing data curation and documentation, (4) developing a better understanding of the data and strengthening privacy and security protections, and (4) expanding data capacity through linkage and storage.

Potential uses of administrative data to generate evidence

Administrative data can help to develop practice-based evidence, which is evidence generated "in the field." Administrative data can provide information that is often difficult for an individual to provide such as laboratory results, specifics of health insurance coverage, or the timing of events. When combined with survey data through linkage or by adding contextual data to a

survey data set, administrative data can bring a previously unavailable richness and depth to an analysis. The data can also facilitate the study of under-researched populations. This benefit of administrative data lies in their ability to fill the gaps in our understanding of social risk factors and at-risk populations that limit the application of evidence-based policymaking to these topics.

Agencies can also use administrative data to assess internal processes and characteristics, such as the grants management process and workforce skills. For example, CDC is using internal administrative data on the software programs installed on employees' computers to gain a new perspective on the skill set of its employees and on the uptake of new software within the agency.

B. Issues

Administrative data have wide-ranging potential for application to policy research and program evaluation. While administrative data can be a rich source of information that can supplement survey data or independently inform program operations, there are a few limitations that researchers should be aware of when applying administrative data for statistical purposes.

Benchmarking the quality of administrative data

Using administrative data for evidence-based policymaking is subject to challenges. Beyond the various legal restrictions on their use, administrative data can also be subject to various degrees of quality and completeness in terms of variables relevant to research. For example, although administrative data may be more accurate for measuring actual financial claims relevant to the program operations, administrative data may be incomplete or inaccurate with respect to participant characteristics such as race.

Curation, documentation, and other analytic limitations of administrative data

Although administrative data hold substantial promise, all types of data have limitations along with their strengths. The AWDG's review of HHS administrative data identified three factors for rating the suitability of the data for research purposes: (1) quality, (2) curation (documentation), and (3) accessibility.[30] Agencies should assess administrative data for its quality on dimensions such as coverage, content, completeness, validity, and accuracy. In addition, the utility of even the highest-quality data as measured by these three factors may be minimal for research purposes if the data are not adequately curated and documented.

Legal and statutory limits on sharing administrative data

Some individual records in administrative data cannot be shared without the consent of the individual. Since the records often contain PII and/or PHI, the legal protections of administrative data must be considered before using them for research. A fuller discussion of privacy concerns is provided in Priority 4.

C. Opportunities

The following two opportunities are proposed to increase the Department's use of administrative data in support of enhancing its evidence-based portfolio.

Opportunity 1: Expand the use of administrative data in the Department

Strategy 1: Improve the documentation and curation of HHS administrative data

HHS could better document and curate administrative data internally to support cross-departmental knowledge and use of administrative data that are available for research. Documentation must be thorough enough that researchers without an in-depth knowledge of a particular data set can still use it confidently and competently based on available information resources. This effort could include developing better metadata for inclusion in the HHS data catalog. The Data Council, via the ADWG, could lead HHS in the development of criteria that define high-value data sets specifically for program evaluation and evidence-based policymaking. The criteria may include factors such as whether the data collection can answer a range of departmental policy questions and/or whether it has useful indicators that can be linked to other data sets to increase its research potential.[31]

Opportunity 2: Improve the quality of administrative data for research

Strategy 1: Develop quality frameworks for administrative data collection

Developing quality frameworks for administrative data collection will help agencies to ensure that their data provide sufficient coverage and content, and that the information is complete, valid, and accurate enough[32] to serve as a reliable source of evidence for policymaking and program evaluation. Currently, the Federal Committee on Statistical Methodology (FCSM) is pursuing quality framework efforts with OMB and the Interagency Committee on Statistical Policy (ICSP).[33] HHS can consult quality frameworks used in the federal government and can also look to other leading frameworks in this space, such as the one used in the European Statistical System (ESS).[34] Building on this work, the Council could take the lead in developing Department-level guidelines on best practices for every step in the data lifecycle of administrative data, including minimum standards for including potentially useful data in a catalog even if the data have not been fully curated. Information about restrictions on access to the data should be included.

Strategy 2: Create procedures to benchmark big data for program evaluation

Administrative data are a major source of big data, but the two terms are not interchangeable, nor is one a direct subset of the other. The complexity and diversity of big data create new challenges in how such data should be evaluated with respect to their performance. Given the increasing availability and use of big data, appropriate evaluation metrics need to be established. Under the auspices of the Data Council, HHS can establish and/or foster procedures to validate and benchmark any new (or newly linked) big data before they are made available to the broader research community.

PRIORITY 3 – INCREASING DATA LINKAGES ACROSS DIVERSE DATA ASSETS

A. Background

The research potential of much of the data that HHS collects can be expanded further by combining them with other data sources through "data linkage." In doing so, HHS can support more complex, informative research and performance monitoring; allow program evaluations to more efficiently track long-term impacts; and provide richer and more diverse sets of outcomes.

Linkage is the process of connecting two or more data sets to enhance the contents of one of the data sets based on a common element, such as a person or business entity. It is most often accomplished through record linkage, defined as any method by which records from different data sources thought to describe the same entity (person, family, neighborhood) are matched[35] The general steps for linking data are to identify the necessary data sets; obtain the required approvals from regulatory authorities, funding sources, and institutional ethics boards; select the data elements that will be used to link the data sets; determine the most appropriate method and matching algorithms for linking, including code systems and terminologies; and assess the quality of the match.[36] Last, the combined analytic file should be assessed for disclosure risk sensitivity; that is, the extent to which the file could disclose confidential information about individuals or organizations. When multiple data sources are combined, the information is enhanced, but the risk of the re-identification of individuals also increases. If achieved well and executed carefully through appropriate methods, data linkages can contribute to the development of evidence-based policy. The types of questions that can be answered through data linkage extend well beyond those that can be addressed through the use of a single data set by combining the strengths that each data set offers.

Advantages to data linkage

Data linkage has several advantages. Person-level data sets can be linked to aggregate data, for example, based either on area of residence, such as county, or on institutions, such as an employer, thus providing the capacity not only to model and control for contextual influences, but also to run complex multi-level models. Linkage of survey and with administrative data can also help to fill in missing values in surveys and to validate responses to surveys, thus reducing the response burden and data collection costs.[37]

In addition to linking survey and administrative data, linking administrative records from multiple departmental programs, other federal and state data collections, and even private enterprises can provide a more complete picture of services received by individuals and families. If these records are also linked over time, they can support predictive models used to identify (1) high-risk populations that may be helped through early intervention and (2) optimal intervention strategies for producing the best long-term outcomes and the largest cost savings. For example, in addressing the opioid crisis, studies have linked data on prescription patterns from state prescription drug monitoring programs or from Medicaid claims with state death certificate data to better understand the relationship between prescribing behavior and overdose death rates.[38] Similarly, federal data could be combined with private sector data on

pharmaceutical sales and nationwide distribution to provide a holistic picture of the prescription opioid supply.

Finally, data from rigorous evaluations of federal programs can also be strengthened by linking the information to state and federal administrative data sources. For instance, program evaluation data linked to earnings, social services, and even death records can enable evaluators to examine long-term impacts of interventions, often at a cost that is far lower than the cost of actively following program participants or developing survey-type evaluation strategies. Linking federal program data to administrative data in randomized controlled trials allows for more efficient data collection and can be a cost-efficient means of following participants' progress over time to assess the medium- and long-term impacts on key outcomes such as work, health, and education. Increasing the use of data linkages in program evaluation will help HHS to more readily achieve its mission by enabling better program evaluation and further research, both of which can improve the health and well-being of the nation.

B. Issues

HHS has made important advances in creating linked data sets and in making them accessible to researchers inside and beyond the federal government. The Department also recognizes the substantial challenges that remain to take full advantage of linking as a tool to unlock the great potential of data that are already being collected. Since this is a complex process, many issues arise in developing successful data linkages, several of which as listed below.

Lack of sufficient documentation

A basis for successful linkages depends on a thorough knowledge of multiple data collections. Such knowledge exists, but it is often limited to the large statistical collections or to the few persons directly responsible for each individual database. In 2017, the Administration for Children and Families (ACF), in partnership with ASPE, cataloged its major administrative databases and surveys, including comparable information on the following linkage-related issues: data linkage capacity; available PII; statutory and practical restrictions on who, what, and where linkage may be allowed; and linkage history. ACF's work should serve as a model for efforts to develop Department-wide data inventories.

Technical coordination

Technical issues must be aligned for data linkages to be successfully conducted. While computer hardware and software issues do cause some impediments, more problematic technical issues reside in the data and the processes that are used in their collection. For example, the interoperability of data that logically could and should be analyzed together may be impossible if the timing of data collections occur on differing schedules, such as on the calendar year, the academic year, or the government fiscal year. Standardization issues with identifiers, such as names, dates of birth, and social security numbers, used to link datasets is another often discussed issue area.

Privacy protections

Data linkage increases the possibility of re-identifying individuals even after the PII has been removed, so access to data must be controlled in order to maintain confidentiality. This often means restricting access to the linked data to secure systems such as the NCHS's RDC system or the Federal Statistical Research Data Center (FSRDC) run by the Census Bureau.

Resource constraints

Although data linkage can be cost-effective compared with primary data collection, it still requires significant resources. The negotiation of the required data-sharing agreements, the time to review and approve research proposals, limited access to secure data systems capable of handling linking and subsequent analyses, and the costs associated with using such systems are all recognized as significant barriers to data linking. Moreover, the organizations sharing the data may have different interests, legal obligations or interpretations, and resource capacities that affect their ability to link the data sets of interest.[39]

C. Opportunities

The following three opportunities are proposed to increase the use of data linkages across a diverse group of data assets in support of enhancing the evidence-based portfolio in HHS.

Opportunity 1: Apply existing departmental knowledge and lessons learned from data linkages

Strategy 1: Develop an HHS Data Linkage Repository that includes information on linkage strategies, barriers, and opportunities

HHS could create a repository of methods for linking data that can help to (1) pool existing departmental knowledge and documentation of best practices, and (2) inform future data linkages. NCHS has documented its past and current data linkages,[40] but there is not a consolidated record of data linkages across HHS at the Department level. Care should be taken to document the following: the process of applying for and receiving access to the data to be linked, the type of linkage, the process for accessing the linked data and the specifics of analysis, appropriate metadata, and major findings and any publications from the linked data sets. This information can expand the knowledge of internal and external researchers about how to approach data linkage for research purposes and prevent the unintended duplication of efforts.

Opportunity 2: Improve the capacity to link HHS data internally and with other data sources

Strategy 1: Promote data linkage between HHS agencies and between HHS and other federal agencies to address departmental priorities

There are two main barriers to linking and analyzing federal data: (1) the lack of processes for how to access linked data for research purposes and (2) the absence of an infrastructure to support more data linkages. Although NCHS has led the Department in linking its data sets to other HHS agency data sets, the demand to link data sets together to derive new insights has

grown, and other agencies have begun to explore the use of storage platforms such as data lakes to facilitate linkages. There are efforts in the Department to better standardize the processes for data linkage, which the Data Council can help to support. For instance, the Council can help to provide technical advice on the form and content of data-sharing agreements for data linkages. This would include tailoring agreements to the unique statutory and regulatory requirements of each major data resource.

Another factor that limits the linking of data is the collection and preservation of identifiers that permit linkage. Several federal administrative data sources could support valuable research efforts, but they lack the necessary identifiers to link to other data sources. In other cases, such data are collected but are then destroyed once the data collection period is complete. Such practices should be reconsidered in an era in which data linkage is being encouraged to increase the value of existing data for research. To the extent practical, this work should build toward federal standards that would function as part of any whole-federal solution like the one envisioned by the CEP's National Secure Data Service (NSDS). The Data Council and the Department can encourage interagency research projects that not only link data sets from multiple agencies, but that also explore data linking with other federal departments to the mutual benefit of all parties.

Strategy 2: Promote data linkage to nonfederal data

The Data Council can strive to expand the use of data linkages to the state and local levels in order to help inform community-level decision making. Encouraging the linkage of HHS data with state and local data would not only support the development of rich, integrated databases; it would also encourage collaboration between the public and the private sectors, opening up access to other valuable sources of data.

The potential for developing rich, integrated databases that are needed to support complex intervention models and their evaluation is far greater at the state level than it is at the federal level, at which necessary data are often missing, and many important data systems (e.g., in education and child care) are not available. The Data Council can look for opportunities to support existing and new efforts to create state and local integrated data systems (IDS), which can then be used to generate data and evidence that are valuable to both participating entities. The support can include waivers, promoting interoperability standards, and technical assistance groups that would help entities interested in further developing such systems.

PRIORITY 4 – MODERNIZING PRIVACY PROTECTIONS

A. Background

The importance of privacy for federal data collection

Collecting data about individuals, facilities, and other entities across the United States is a key part of HHS activities. Increased access to data for research and program evaluation has the potential to produce greater knowledge and innovations that will enable HHS and other parts of the federal government, as well as state and local governments, to make informed policy and

program decisions. However, increasing access to and the use of HHS data requires understanding the legislative landscape, assessing the privacy risks of collecting certain data, and carefully applying the procedures and protocols designed to protect the privacy and confidentiality of individual information that the public provides to HHS.

Agencies across the federal government, including HHS, have a special obligation to establish and maintain the public trust and to satisfy the public's expectation that data provided to the government will be used responsibly and protected from inappropriate disclosures. HHS OpDivs and StaffDivs often collect data that contain highly sensitive information about vulnerable populations that sometimes include PII; this information should never be shared beyond the authorized individuals working with the data. Inappropriate disclosure of these data can lead to a breach of trust with the public, which can cause people or entities to not respond to requests for information or to submit incomplete or inaccurate data. Without accurate data from the public, HHS cannot effectively carry out its mission to protect the health and well-being of the American people.

The government's assurance that the public's data will remain private is absolutely critical to maintaining public trust and thus ensuring that HHS has accurate information to inform policy and program decisions. Gaining appropriate consent from individuals to collect their information grants individuals more control over how information about themselves will be used. In order to maintain the trust that the public places in us and to ensure that the public continues to allow HHS to collect personal information, it is incumbent upon data custodians at HHS to be transparent about their intentions and to manage data in accordance with (1) the promises that HHS makes to data providers and respondents and (2) with privacy laws and requirements.

There are a variety of policies and laws that protect the privacy of individuals' data and that apply to HHS. A fuller discussion of these policies and laws appears in Appendix B: U.S. Privacy Policies and Laws Applicable to HHS.

B. Issues

There are many laws and regulations in place to protect the privacy of individual data, and HHS must continue to protect the privacy and confidentiality of its data in the current environment. This is made increasingly difficult by the growing availability of data in the public domain. The Department should remain forward-thinking in the ways that current privacy law is applied to data used for research purposes, and it should develop effective protocols to prevent unauthorized disclosure of data. Some of the issues and challenges raise by the need to protect the privacy of individual data are discussed below.

The risk of re-identification

One challenge associated with modernizing privacy protections is keeping regulations, guidance, and disclosure practices up to date. In theory, the accelerated rate at which the release of "de-identified" data files, or PUFs, are being made available could increase the risk of re-identification. Although a data file, if evaluated on its own, may satisfy a standard of de-identification for public release, it is possible that, if the file is combined with other publicly

available data sources, the combined data would be more likely to allow re-identification. This concept has been described as the "mosaic effect."[41]

Additionally, it is not just individuals, but also facilities and other entities that are at risk because of the mosaic effect. It is important for the Department to ensure that privacy and confidentiality protections keep up with the increasing volume of available data and with the technological advances that facilitate sophisticated analysis and re-identification.

The complex matrix of privacy laws

Another challenge is the matrix of privacy laws that govern data use and disclosure that must be considered in making data available for research. In some circumstances, multiple laws and regulations may apply to the use of a certain data set. The overlapping coverage of multiple laws may lead to confusion among those who handle data. It is especially important for guidance at the Departmental level to inform staff of whether data can be shared and the purposes for which it may be shared when applicable.

C. Opportunities

The following two opportunities are proposed to achieve two goals: to modernize privacy and security protections in support of enhancing the evidence-based portfolio in HHS and to help minimize the threat of privacy breaches, including the three issues described in the previous section.

Opportunity 1: Increase data sharing without eroding privacy protections through better communication and coordination with experts

Strategy 1: Use data intermediaries to facilitate HHS data sharing in accordance with privacy laws

Given the wide range of HHS data assets and variation in agency data-sharing policies, including legislative requirements, the Data Council could recommend approaches to identifying the roles and functions of a data intermediary who would help to coordinate data-sharing policies across HHS agencies. The intermediary could also help to navigate the framework of laws and provide guidance on best practices for data sharing at the departmental level.

Strategy 2: Increase the use of disclosure review boards or data disclosure boards and provide guidance on best practices for de-identification

A number of HHS agencies use disclosure review boards or data disclosure boards to ensure that data sets are released in a way that complies with all applicable federal laws, regulations, and agency polices, including preventing the unauthorized disclosure of protected information (e.g., PII and PHI). However, there is variability across and even within agencies in the way that data are masked and de-identified to create PUFs. This variation in practices can lead to an increase in the risk of unintentional re-identification if data are combined, even when appropriate precautions are taken. The Data Council could coordinate with disclosure review and data disclosure boards to develop recommendations on Department-wide policies that have the potential not only to increase access to agency-specific data assets but also to strengthen the

privacy and security protections of those data sets. The Council could also develop best practices for preparing PUFs. Guidance for agencies on best practices for de-identification and for when de-identified data sets can permissibly be shared will help to align departmental processes and ensure that PUFs pose the lowest possible risk of re-identification as a result of the mosaic effect noted earlier. Additionally, the Data Council could oversee the development of continuing education modules for new review board members and updates for existing members.

Opportunity 2: Assist in standardization of departmental privacy policy practices

Strategy 1: Support the Department's development of more streamlined data sharing processes, including model enterprise-wide DUAs and inter-agency agreements (IAA)

As part of the Agency Priority Goal of combining disparate data sets, there is an effort to develop an enterprise-wide DUA.[42] Enterprise-wide DUAs should also reflect differences across agencies in requirements for sharing of data. The Data Council can coordinate the development of issues that DUAs should address and provide standard language that would apply in most cases. The Council would also focus on the need to include appropriate privacy and security protections in the DUA based on the type of data it covers, with whom the data are being shared, and the purpose for which the data are being used. Implementing and promoting components of a model DUA will help to better standardize and align departmental policy such that it informs data sharing while protecting privacy.

In addition, the Data Council can review the current IAA process and documents and make recommendations on how to streamline the process. Implementing a more streamlined IAA process will accelerate the agencies' ability to share data.

Strategy 2: Address privacy and legal concerns about the use of data for policy research, statistical purposes, and program evaluation

HHS agencies should work toward developing a systematic approach for examining data sets and for assigning a level of potential re-identification to them. In an effort to increase the use of administrative and other specific data for research and policy (as similarly called for in Priority 2), agency staff will need to (1) thoroughly examine existing PII and PHI, along with the process for collecting these data, (2) review existing legal restrictions, (3) provide guidance on the appropriate use of administrative and other specific data collections for research and program evaluation, and (4) establish a clear policy on access for researchers seeking to use the data for policy research or program evaluation. Information on whom to consult for privacy questions at various agencies in HHS should be cataloged wherever possible and included with the information on each data set.

PRIORITY 5 – INCREASING DATA POLICY COORDINATION AND INFORMATION SHARING ACROSS THE DEPARTMENT

A. Background

Within HHS, there are many different stakeholders who share roles and responsibilities regarding data. At the Office of the Secretary (OS) level, these stakeholders include the Division

of Data Policy within ASPE (DP), the Office of the National Coordinator for IT (ONC), the Office of the Chief Technology Officer (CTO), and the Office of the Chief Information Officer (OCIO), to name a few. HHS also has a principal statistical agency, NCHS, which is part of CDC, and a federal statistical unit, the Center for Behavioral Health Statistics Quality (CBHSQ), which is part of the Substance Abuse and Mental Health Services Administration (SAMHSA). Both are authorized by statute and recognized by OMB as having a specific mandate to develop official, nonpartisan statistical information.[43] Beyond NCHS and CBHSQ, HHS agencies collect data for both statistical and research purposes, and administrative data that may be used for secondary purposes. Furthermore, each OpDiv plays a role in shaping data and statistical policy for its own agencies. Finally, there are Federal Advisory Committees which help formally bring in outside perspectives to the Department on policy matters. Although each of these groups provides guidance on data policy, there is no formal mechanism for ensuring coordination across the groups. As a result, there are overlapping roles and gaps in managing the data. A better understanding of these groups, their efforts, and the implementation of some overarching role should help to mitigate these issues. The following discussion describes some the major data policy actors in OS and HHS. Descriptions of the offices and links to their specific websites can be found on the HHS website.[44]

Office of the Secretary

Individual offices

The **DP**, which is part of **ASPE**, coordinates departmental data and statistical policy. These activities include conducting evaluation and policy research to increase either data capacity or the efficiency of current data collections, sponsoring data collection modules to improve data quality or to provide for methods development and testing, and understanding the various data collections' strengths and limitations in terms of informing policymaking.

CTO fosters innovation across HHS. This includes cultivating public-private collaborations to improve government efficiency, working toward making HHS data available to the public, and providing technological solutions to improve HHS operations.

OCIO, which is part of the **Assistant Secretary for Administration (ASA)**, is the main office for ensuring that HHS follows good business and security practices in implementing IT. This work includes inventorying IT investments, monitoring the overall HHS IT program to maximize efficiency, implementing cybersecurity and privacy provisions, and supporting the OpDivs with their business and IT system investments.

ONC supports the national adoption of health IT and the promotion of nationwide health information exchanges to improve care. This work includes establishing data standards that support the interoperable exchange of patient records, developing and implementing procedures for certifying EHRs, and implementing specific provisions of the 21st Century Cures Act.
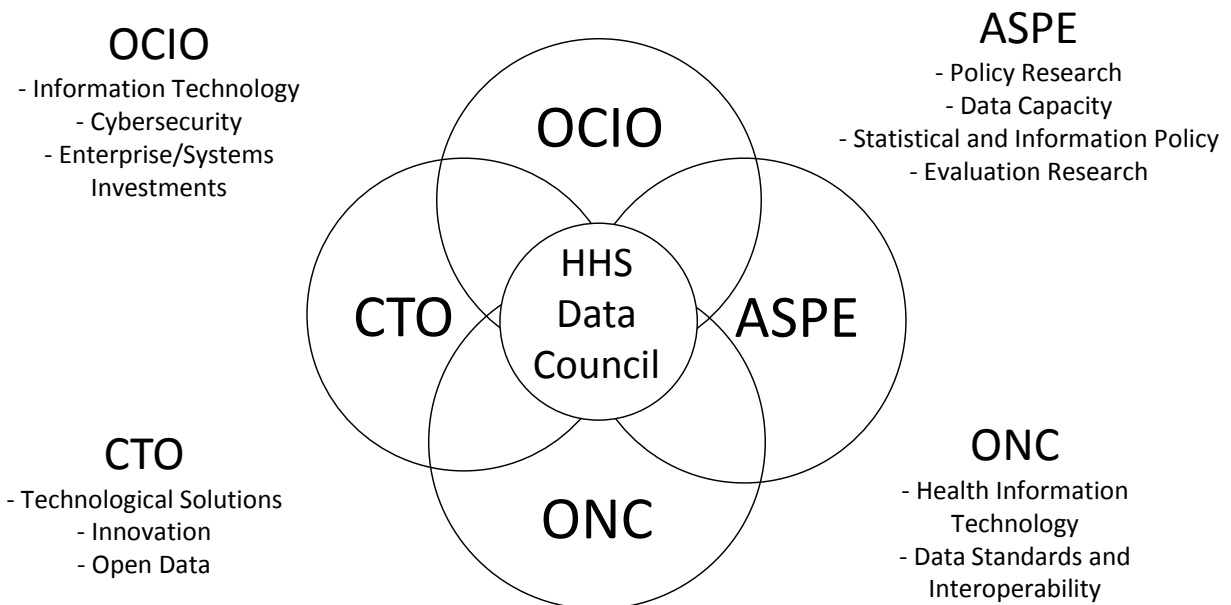
OS-level councils

The **HHS Data Council** advises the HHS Secretary and provides leadership to the Department on health and human services data and statistical policy. When the HHS Data Council was

established in 1996, its initial mandate was to coordinate data collection activities and statistical policies across OpDivs and StaffDivs; meet the data needs of all users; reach a consensus on data standards and privacy; provide oversight of surveys and statistical analyses; and lead an interagency effort on data standards, privacy, telemedicine, and enhanced health information for consumers. While the Council continues to serve as the leading body on data and statistical policy, its mandate and role in policy coordination has evolved over time as ONC and OCIO were established to manage the Department's information policy.

The **Enterprise Risk Management (ERM) Council** comprises senior leaders in the Office of the Secretary and Chief Operating Officers from OpDivs who coordinate ERM across the Department. Per OMB Circular A-123, Management's Responsibility for Enterprise Risk Management and Internal Control,[45] the ERM Council oversees the development of Department-wide Risk Profiles, which are an internal management tool for identifying, assessing, and strategically prioritizing HHS's most significant risks and opportunities. The ERM Council assess risk across a wide variety of financial, administrative, strategic, and mission-critical elements. As the Department makes decisions to (1) implement enterprise data governance, (2) upgrade IT systems to better protect information security and the privacy of HHS data, and (3) account for a changing budget and policy environment, it will be important to consult the ERM Council and account for risk in policy and programmatic decision making.

The main data stakeholders in the Office of the Secretary and their primary policy functions are shown in Figure 1.

Figure 1: Data Stakeholders in the Office of the Secretary, HHS

OCIO
- Information Technology
- Cybersecurity
- Enterprise/Systems Investments

ASPE
- Policy Research
- Data Capacity
- Statistical and Information Policy
- Evaluation Research

CTO
- Technological Solutions
- Innovation
- Open Data

ONC
- Health Information Technology
- Data Standards and Interoperability

OCIO
CTO
HHS Data Council
ASPE
ONC

Federal Statistical Agency and Recognized Statistical Unit

NCHS is the nation's principal federal agency for health statistics. As such, NCHS abides by the principles and practices set forth by OMB to produce reliable, credible, unbiased, accurate, and timely health statistics.[46] For nearly 60 years, NCHS has conducted core data collection activities across a full range of health conditions and issues related to health care. The agency obtains and disseminates data from vital records, population-based surveys, surveys of providers, and administrative health care and other records. These data are used to describe and monitor health conditions, risk factors, and health care utilization. The Center also provides data that other data systems use as a benchmark. NCHS also has played, and continues to play, a key role in setting data-related standards for the Department.

Housed within SAMHSA, **the Center for Behavioral Health Statistics and Quality (CBHSQ)** is the Department's lead agency for behavioral health statistics. As a federal statistical unit, CBHSQ abides by the same principles and practices that govern NCHS and other federal statistical entities. CBHSQ ensures that data collection, analytic activities, dissemination activities, and evaluation efforts are consistent with the mission and priorities of the Department and SAMHSA. It also participates with other federal agencies in developing national policy on statistical matters. CBHSQ oversees several data collections, disseminates reports and data to the public, and advises the Assistant Secretary for Mental Health and Substance Use and the Department on behavioral health data and statistics.

OpDivs

HHS is one of the largest departments in the federal government. This makes it much more important and inherently difficult to coordinate all the agencies that make up HHS. Each OpDiv houses its own groups that coordinate its intra-agency efforts on data policy. It would be beyond the scope of this report to detail the role each OpDiv plays in forming data and statistical policy. This section provides a few examples from five OpDivs. In January 2007, NIH established the Council of Councils, which makes recommendations for research on emerging scientific opportunities, rising public health challenges, and gaps in knowledge. In addition, the National Library of Medicine (NLM), an institute in NIH, is the central coordinating body for clinical terminology standards in HHS.

The efforts and institutes in the CDC and ACF provide additional examples of where OpDivs play a role in forming data and statistical policy. The CDC launched a Surveillance Strategy in 2014, which was designed to improve the availability and timeliness of data; advance the use of EHRs, mobile technologies, and cloud computing; reduce reporting burden; and retire redundant systems in health departments. ACF's Office of Planning, Research, and Evaluation (OPRE) plays a coordinating role in human services research, including data analysis, evaluation, and performance management. Finally, the Agency for Healthcare Research and Quality (AHRQ) and the Centers for Medicare and Medicaid Services (CMS) have advisory councils and internal bodies that provide guidance on data policy. In addition to the above examples, there are plenty of additional governing bodies throughout HHS that facilitate data policy coordination.

Although each OpDiv and StaffDiv has a specific mission, many health and human service issues span several sectors and require the combined effort of more than one agency. This is where coordinating groups, like the HHS Data Council, provide an opportunity and a forum in which to unite the efforts and entities within each agency at the Department level. Data Council meeting agendas are shaped by the membership of the Council and thus reflect issues that affect more than one agency. The diverse representation on the Data Council allows it to address data priorities for the whole Department. Figure 2 shows the relationship between the Data Council and the various agency and OpDiv stakeholders represented on the Council.

Federal Advisory Committees

Federal Advisory Committees can provide an important connection between the public and the Department, and are formed under rules of the Federal Advisory Committee Act.   A committee relevant to data policy is the National Committee on Vital and Health Statistics (NCHVS). The NCVHS states it "serves as the statutory [42 U.S.C. 242k(k)] public advisory body to the Secretary of Health and Human Services for health data, statistics, privacy, and national health information policy and the Health Insurance Portability and Accountability Act (HIPAA). The Committee advises the HHS Secretary, reports regularly to Congress on HIPAA implementation, and serves as a forum for interaction between HHS and interested private sector groups on a range of health data issues." The NCVHS has staff that participate on the Data Council, and the Data Council has frequently advised and collaborated with NCVHS on improving the quality of vital statistics and privacy and information standards.

Figure 2: HHS The Data Council's Policy Coordination
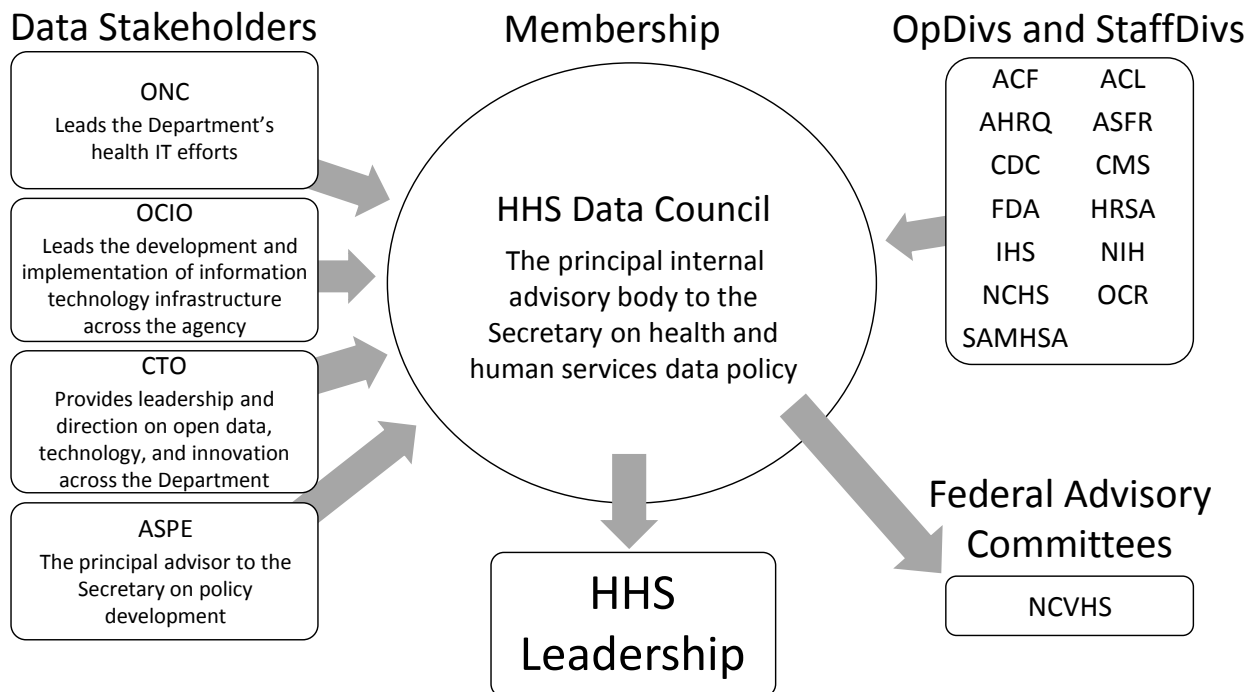
B. Issues

The Department has a wealth of knowledge and expertise in the use of data for evidence-based policymaking. However, this knowledge often remains in siloes, impeding what would otherwise be collaboration on high-priority data needs as well as the consistent and methodologically sound application of HHS data to policymaking. This issue is described below, along with others that prevent better policy coordination in the Department.

Decentralized data collection and management in HHS

The Department is a complex collection of OpDivs and StaffDivs that both act independently and have a diversity of missions. The decentralized nature of the data collection system within HHS and in the federal government overall result in data systems that reflect programmatic needs but that also create an inter-agency culture of independence. Data collection and management processes are usually at the OpDiv level, where data collections originate and are managed by individuals who are most familiar with the information. However, the unintended result of this decentralization is the loss of opportunities to coordinate and benefit from data collections across HHS.

Limited coordination between OCIO, CTO, ASPE, and the HHS Data Council

Although there are established bodies at the OS level to coordinate Department-wide data policy (the HHS Data Council, the CIO Council, and the CTO), the breadth and depth of each of their substantive areas leaves little time to communicate and strategize together. Additionally, these entities may not have a clear path of coordination when it comes to data policy, information policy, data curation, and data management. Limited coordination between relevant parties results in less efficient processes and an increase in the risk for the unintended duplication of efforts.

Communicating the value of HHS data to policymakers

HHS collects a wide variety of data that are relevant to policy and program decisions; however, there is often a disconnect between the staff who curate, manage, and analyze HHS data, and the policymakers who make decisions for the Department with respect to the value and purpose of HHS data collections. Communicating the research and policy value of HHS data collections from survey to administrative data is critical to gaining buy-in from within HHS to use the data for evidence-based policymaking. Furthermore, communicating the value of HHS data collections is important in securing funding for the continual improvement and maintenance of these efforts.

C. Opportunities

The following two opportunities are proposed to improve data policy coordination and information sharing in the Department in support of enhancing its evidence-based portfolio.

Opportunity 1: Increase coordination in the Department regarding data collection, system and software investments, and data management and governance

Strategy 1: Identify high-priority data and information policy issues that require OS-level coordination

As mentioned, there is no single entity that coordinates all of the different aspects of HHS data. Although there are various councils, including the Data Council, the CIO Council, and the CTO Council, these bodies do not always coordinate with each other to recommend or create policy. Identifying high-priority issues on which these councils can coordinate and developing mechanisms for coordination between the different governing bodies would cultivate a unified approach to high-level data and information policy issues that affect the entire Department with respect to data collection, management, and release. As the Data Council has broad membership, it could act as the convening agency to identify the high-priority data and information policy issues requiring OS coordination.

One area of data policy that has already been identified as a high-priority area for the Department to coordinate is data governance, which is defined as a set of processes that ensure that data assets are formally managed such that departmental needs are met. A data governance model such as this would establish authority and define the parameters for the data produced or managed by the enterprise. Developing an HHS-wide data governance plan was named an Agency Priority Goal (APG) for FY2018 as a way to improve interagency data sharing and policy coordination.[47] The Data Council could serve as the convening body to support the Department's efforts in the development of a governance plan, as described in the Agency Priority Goal, including the identification of the key items that need to be included.

Strategy 2: Increase communication between the Data Council, OCIO, ONC, and the CTO

The HHS Data Council recognizes that DP, ONC, CTO, and OCIO each have a distinct role in data collection, data usage, data protection, and data dissemination across the Department. OCIO[48] focuses primarily on the business and infrastructure components of a data collection system, DP[49] focuses on the policy implications and decision-making potential, ONC[50] leads data efforts related to health IT and interoperability, and CTO[51] provides leadership on technology, innovation, and data sharing. The Council can work to increase communication and collaboration in areas where two or more offices contribute to policy formation. Data Council meetings can also be used as a forum for seeking guidance on or assistance with particularly challenging issues. Building on existing forums and councils, and using the Data Council as a source of expert guidance, HHS can encourage collaboration and strengthen its data resources.

Opportunity 2: Inform policymakers and researchers about the value and uses of HHS data

Strategy 1: Communicate the value of HHS data collections and systems for policymaking

It is critical for HHS data stewards to communicate the content and potential uses of the Department's data as a valuable resource in the effort to inform policy in a more coordinated fashion. HHS data are a public good in that they are used for decision making that ensures that

programs and policies are serving the needs of the American people. The data collection systems supported by HHS provide most of the capacity to monitor the health of the population and the functioning of the health care and human services systems. Data resources are essential to identifying national trends and to informing program and policy decision making associated with the missions of all HHS agencies. Each HHS agency data collection fills a unique gap and captures information needed for a specific purpose.

The Data Council and HHS agencies should work toward improving ways to promote the value of HHS data through, for example, return-on-investment reports, to those who may influence future funding. Maintaining a high level of quality and a wide scope for HHS data collections requires continual budgetary investment. HHS data stewards need to actively communicate the value of HHS data to policymaking as well as the investments necessary to maintain their usefulness to further improve policy and program decisions.

Strategy 2: Inform internal stakeholders about potential tradeoffs between timeliness and quality and how this impacts the usefulness of data for policy purposes

In the current computing environment, the ability to collect, process, clean, and disseminate large, population-based data has enabled agencies to substantially speed up data releases. Early data release platforms, such as those used by programs like NCHS's National Health Interview Survey Early Release program, allow data that have a five- to six-month lag time from initial collection to generating provisional estimates on a quarterly basis.[52] Mortality data on drug-related deaths from the National Vital Statistics System (NVSS) are also now released on a monthly basis with an approximate six-month lag from the date of event on a data visualization platform.[53] For data systems across HHS, there have been significant improvements from previous release schedules for many types of data.

Although faster release of data is a highly sought-after and valued goal, agencies need to ensure that they continue to achieve their missions by producing reliable, high-quality data regardless of the timeliness or release mechanism. Data that are disseminated by early release programs are almost always provisional or preliminary, not the final data file for a given period. It is therefore imperative to ensure that proper processing and cleaning techniques are still used. Depending on the size of the data collection, there may be a tipping point at which data cannot be released any faster while still maintaining quality and comparability to final data. Expectations of the Department and the agencies collecting the data should be realistic and well-informed. The agencies are responsible for educating Department leadership about the feasibility, limitations, and ramifications of earlier data releases, whereas the Department is responsible for maintaining realistic expectations of the limits of data and agency resources. It is also important for data curators to communicate to policymakers a realistic timeline for data collection and release that will maintain the data quality.

PRIORITY 6 – BUILDING A 21ST CENTURY DATA-ORIENTED WORKFORCE

A. Background

Data-oriented workforce

Statisticians, informaticians, computer scientists, data scientists, and other similar professionals across the Department are involved in statistical and other data analyses that meet programmatic and research needs. There is a growing recognition that we may be reaching a workforce crisis as these experts retire and there is greater competition in hiring staff to replace them. At the same time, there is also the need to develop new skills in existing staff and to hire new staff with a wider range of skills. Promoting the retention of the existing statistical staff and the broader data-related workforce across HHS, and providing opportunities and incentives for staff to continue to develop their skills to meet the needs of an evolving data and statistical environment are important, necessary goals. The creation of this 21st Century data-oriented workforce must be done in the context of the evolving field of data science.

Data science

Over the last decade, a huge and ever-increasing amount of information has been generated and digitized from an increasingly wide variety of sources. That information has been combined with new technologies and analytic methods to offer new opportunities to develop new insight into our world. The sheer volume and diversity of this growing information, the complexity of the underlying technology, and the increasing variety of analytic methods have presented technical and cognitive challenges to making sense of available data. The presence of such data, technology, and analytic methods are also creating a demand for staff who can extract, mine, and analyze huge and complex datasets, and then explain what they mean to the public and decision makers. In response to these challenges, the new field of data science has emerged. The National Academies of Sciences, Engineering, and Medicine have described data science as follows:

> [Data science centers] on the notion of multidisciplinary and interdisciplinary approaches to extracting knowledge or insights from large quantities of complex data for use in a broad range of applications. Data science is about synthesizing the most foundational disciplines to solve particular classes of problems or applications that are newly enabled because the volume and variety of data available are expanding swiftly, data are more available immediately, and decisions based on data are increasingly automated and in real time.[54]

One of the pillars of the current President's Management Agenda is building a workforce of the 21st century, which includes empowering leaders and managers to align staff skills with the evolving mandate and responsibilities for which federal agencies are expected to deliver.[55] Similarly, among the recommendations of the CEP was a call to give federal departments the authority to construct a high-performing, evidence-building workforce.[56] Doing both at the same time will require the Department to expand and leverage a data science workforce.[57] To effectively develop evidence that can inform policymaking and program evaluation, HHS could expand the analytical and computing capabilities of its staff so that it can work with the large and

complex data sets that are increasingly being generated or accessed by and for the Department's programs.

B. Issues

HHS expertise has not kept pace with recent innovations in the use of data to analyze problems and support decision making, including the application of data science. While there is interest in integrating data science into HHS programmatic operations and policy research, there are administrative and resource challenges that have limited its application. Furthermore, as the field of data science continues to evolve, there are new challenges regarding how to meaningfully incorporate such data analyses into policymaking in a methodologically sound way. The following sections discuss a few of the obstacles to building a data-oriented workforce that meets 21st century needs.

Departmental staffing limitations

Mastering and applying data science requires a multidisciplinary skill set, including an understanding of math and statistics, computer science and programing, communication and visualization, and subject expertise in the specific field of interest being studied. Many individuals with data science skills are trained in disciplines from which HHS does not typically hire, including engineering and the physical sciences. HHS must maintain its current statistical workforce, prepare other Department personnel for an increase in work related to data science, and expand the work force to provide additional expertise in data science and its related fields. However, it may be difficult to recruit and retain enough data scientists through the existing federal hiring and promotion processes. For example, current processes hinder HHS from being competitive with the private sector, which offers more lucrative jobs and more flexible and streamlined hiring processes. These barriers must be overcome in order to gain access to the full talent pool of data scientists.

The need for expertise in interpreting complex data

The availability of more data for analyses does not necessarily translate into better analyses or more robust conclusions. The same scientific rigor in study design and implementation that is required of smaller statistical studies is required when analyzing large, complex, and alternative data to draw appropriate inferences from the results. Invalid conclusions can happen for a number of reasons. For instance, appropriate analyses based on large data sets require a fairly high level of training, complex sampling designs require results to be weighted, and very large samples can easily produce statistically significant results that are not substantive or meaningful in practice. The increasing availability of data is not a replacement for a substantive understanding of a given field. A data analyst without enough in-depth subject matter expertise on a topic may view a statistically significant result as unique or important, in contrast to a subject matter expert who would view the same result as spurious or irrelevant. It is important to understand the valuable contributions that data can make to policy decisions, and it is equally important to understand the limitations of particularly large, complex, and alternative data sources, such as those obtained from the private sector, sensors, or websites and social media.

Data ethics

As the volume and use of data have grown exponentially, the risks associated with analyses of these data have also increased. As noted in Priority 4, one risk associated with increased data collection and analyses is an increase in the potential for losing individual and group privacy. Beyond privacy, there is the increasing risk that inappropriate research designs or methods will be applied to data analyses, leading to illegitimate or biased results and conclusions. Furthermore, as there are greater and greater volumes of data to potentially use and more individuals using it, there is an increase in the risk that data will be selectively chosen or even manipulated to leverage the interests of an individual or group over those of the rest of the general population. Consequently, data ethics should be consistently and holistically considered and applied. Although a universally accepted definition of data ethics has not been created, one reasonable summary describes data ethics as follows:

> A new branch of ethics that studies and evaluates moral problems related to data (including generation, recording, curation, processing, dissemination, sharing and use), algorithms (including artificial intelligence, artificial agents, machine learning and robots) and corresponding practices (including responsible innovation, programming, hacking and professional codes), in order to formulate and support morally good solutions (e.g., right conducts or right values).[58]

Other groups, such as the Council for Big Data, Ethics, and Society, have suggested rules to guide researchers through the realm of ethics in large and complex data. Such rules include acknowledging that data can do harm, developing codes of conduct, and engaging in the broader consequences of data and analysis practices. These types of rules provide an idea of the types of issues that must be addressed.[59]

Technology and systems needs

An increase in the Department's volume, use, and sharing of data requires the necessary technology and systems to handle data science demands and to secure data against information breaches and the inappropriate disclosure of PII and PHI. HHS has already developed systems to help address these issues. The RDCs, for example, provide a secure physical environment in which data can be accessed. In some cases, researchers can also download aggregated reports and results to their own personal workstations. In addition and where permitted, a number of software programs or environments allow researchers to run queries against data files, but they receive only de-identified aggregate results. However, these systems are still limited in capacity and flexibility, which can interfere with a data scientist's ability to use available analytic techniques. In addition, proposed modern technology solutions, such as the NSDS, could help to defuse inferential disclosure and provide data encryption and storage to counter security threats if implemented. Such new technologies should be evaluated for their suitability in data science and for their ability to secure PII and PHI.

C. Opportunities

The following opportunities are proposed to build a 21st century data-oriented workforce in support of enhancing the evidence-based portfolio in HHS. These opportunities are important because the evolution of data science is part of what is creating the push behind and the connections between the first five priorities and their associated opportunities.

Opportunity 1: Enhance the data science capability of the current HHS data workforce

Strategy 1: Increase data science and statistical training opportunities

HHS can both provide a higher level of internal training opportunities related to data science and statistics and raise the staff's awareness of these opportunities. Several HHS agencies have already developed training programs in data science methods and in the basics of HHS data. These programs can be continued, expanded, and better promoted. The need for training in data science and statistics is driven on multiple fronts. For example, HHS analysts and statisticians currently clean, prepare, and transform administrative data for linkage and analysis. However, there is an opportunity to bring in additional personnel who are more adept at data manipulation but may not yet have the advanced training needed for activities such as complex statistical analyses or deterministic/probabilistic data linkages. Similarly, by training staff in data visualization skills, HHS could more effectively present research findings in graphs and visuals that are more understandable to the public. Additionally, many data sources that are potentially useful to data scientists are also large and complex, requiring new methods of analysis, such as artificial intelligence and machine learning. To apply these methods to such sources requires a high level of computing expertise. Training will help to augment staff expertise and prevent issues associated with interpreting complex data accurately. However, it should be noted that additional training of staff is not sufficient to make full use of applying data science to HHS research and program evaluation. Investment in modernizing technology is a necessary corollary to ensure that staff can apply their knowledge of data science effectively. Both the necessary training and associated modern technology are required to first use and then optimize HHS capabilities in data science.

Strategy 2: Promote multidisciplinary data science teams and increased cross-program collaboration

HHS can promote multidisciplinary data science teams and increased cross-program collaboration. Data science involves a wide variety of analytical and programming skills. It also requires some degree of expertise in the field to which data science is being applied, in this case health and human services, to avoid drawing inappropriate conclusions. Few people have all of the necessary skills to effectively apply data science in any given circumstance. Building multidisciplinary teams and increasing cross-program collaboration would allow HHS to leverage different areas of expertise in decision making. This approach would also help the Department to keep pace with an ever-evolving landscape of analytic tools. Parts of the Department are already using this approach successfully. For instance, the Office of the Inspector General (OIG) has built a data science team to uncover billions of dollars in fraud, waste, and abuse in Medicare and Medicaid.[60]

Strategy 3: Promote awareness and education of data ethics in the Department

As HHS trains and expands a data-literate workforce, the discussion of data ethics should play an important role in that effort in order to continue to mitigate data risks, maximize the value of data, and protect the trust that Americans have in the Department to collect, hold, and use their data appropriately. HHS can work to develop guidance and foster a culture of data ethics both within the Department and across the health and human services landscape. The Data Council can help to coordinate policy between relevant stakeholders in the Department to provide the foundation for guidance and education on data ethics.

Opportunity 2: Reinforce capacity to explore the application of data science and alternative data to HHS research and program evaluation

Strategy 1: Develop capacity to investigate new or more-blended statistics for health and human services

HHS is working at the forefront of data science in many respects, but there is still room for improvement. For example, departmental data scientists are bending, stretching, and blending statistical and computational techniques to increase their ability to manipulate and analyze greater volumes, varieties, velocities, and variabilities of and within data.[61] HHS can support the workforce in investigating whether there is a need for such new or more-blended statistics relating to health and human services; the department can also allow more programs to hire employees with these capabilities. Much of traditional statistics is based on sample size, in which a larger unbiased sample translates into more confidence in the research findings. The same methodological rigor should be applied to evaluating large and complex data as is used for evaluating more traditional data sets such as survey data. However, the application of traditional statistical methods to complex data is making it increasingly difficult to tease out actual effects from statistical noise. There is a need for staff who can further examine the use of effect sizes or other indicators of significance that do not rely on sample size for health and human services policy, program evaluation, and population monitoring. Trained staff will be needed to create guidance for evaluating whether statistical differences observed when analyzing complex data are substantively meaningful to the research or program being addressed.

Strategy 2: Develop the capacity to coordinate the evaluation of alternative data sources

The important use of administrative data as an alternative data source has already been discussed at length in Priority 2 of this data strategy. However, administrative data are not the only alternative data source that has emerged as a viable opportunity for data scientists. Other alternative data sources of potential benefit to the health and human services field include, but are not limited to, private sector e-prescription and consumer purchasing data, environmental and health sensor data, and social media data.[62] HHS could ensure that the workforce has the expertise to coordinate the evaluation of these alternative sources of data to determine whether they may be useful for a variety of statistical, research, and programmatic purposes across the Department.

Strategy 3: Explore the NIH Strategic Plan for Data Science as a tool to support data science across the Department

Although the NIH Strategic Plan for Data Science[63] applies to a more specific array of scientific data than are considered in this plan (e.g., data from neuroimaging, genomics, and animal research), its outlined goals and actions provide a blueprint for improving the data science infrastructure across the Department. For example, the first two goals in the NIH plan are to (1) support a highly efficient and effective biomedical research data infrastructure and (2) promote modernization of the data-resources ecosystem. NIH is looking to accomplish the first goal by optimizing data storage and security and by connecting its many data systems. To achieve the second goal, NIH plans to modernize its data repository ecosystem and support the storage and sharing of individual data sets. The Department could explore opportunities to build on NIH's work and to potentially use it as a basis of planning for future data science efforts, as appropriate.

Opportunity 3: Invest in the future of data science

Strategy 1: Increase the number of new data scientists

Beyond enhancing the data science capability of the current HHS data workforce and exploring the application of data science to HHS research and program evaluation, there is a need for more new data scientists within the Department. Coming from budding data science programs or the private sector data science industry, these scientists will bring with them new ideas, techniques, and skills that can both complement and improve existing HHS systems. To reach this goal, HHS and the federal government as a whole could invest in new human resources (HR) strategies not only to encourage these individuals to join the public sector but also to facilitate the onboarding of these individuals in a timelier manner.

The Department can use numerous strategies to successfully hire data scientists for the federal workforce. Four approaches stand out. First, the Department can better promote the benefits of federal service at job fairs and industry events that target data scientists. Second, the Department can explore opportunities for creating new intern and fellowship programs tailored to data science, seeking out potentially qualified job candidates from industries and disciplines not traditionally associated with HHS (e.g., engineering, physics, and computer science). Third, the Department can also place more individuals from these non-traditional occupations in more senior positions. Fourth, to benefit from the broadest data science talent pool, HHS can explore opportunities for making better use of existing specialized hiring programs, like the Presidential Management Fellows program, the Pathways program, the U.S. Digital Service, and Title 42. The Department can also advocate for creating similar programs aimed directly at data scientists with a STEM background.

Strategy 2: Ensure that staff have the expertise to explore the coordinated implementation of technology or software that facilitates ethical data sharing and use for data science capabilities

Data science is being driven by the increasing ability to access diverse data sources. In recent years, there has been an influx of innovative technologies and software programs that facilitate data access (as sought after in Priority 1) while protecting privacy (as sought after in Priority 4).

For example, certain types of software now allows remote access to data housed in data archives. However, as technology evolves and the need to link diverse data sets grows (see Priority 3), the Department must continue to innovate its approach to securing and housing data from multiple sources. Through exploring technology or software that facilitates data sharing, HHS will be better prepared to facilitate research and interdepartmental collaboration while increasing the privacy and security protections for its data assets. This investment must be in technology and in the people who handle technology and data on a daily basis, as recommended in the NAS report.[64] The Data Council, in consultation with OCIO and CTO, could provide guidance on how to evaluate the appropriateness of new or modified technological approaches or software programs to (1) increase ethical data sharing and use for data science purposes across the Department and (2) ensure that appropriately trained staff are available to take advantage of the new technology.

CONCLUSION

With its expansive volume of data assets, HHS has the potential to use its data resources to answer some of the most pressing policy questions of our time, including stemming the opioid epidemic, lowering prescription drug prices and the cost of health insurance, and improving health outcomes for vulnerable populations. Far too often, HHS data resources are underused for six reasons: (1) the data resources available for research are not documented Department-wide; (2) the data are not curated in a standardized way across their lifecycle; (3) current funding streams for data systems do not allow for innovation and puts the cost burden on the researcher; (4) there are legal and agency cultural barriers to sharing data resources and to linking data sets for research; (5) there is limited departmental expertise in using multiple types of data for policy research and analysis; and (6) current computing systems and cybersecurity need to be improved to protect the privacy of individual data and facilitate the analysis of multiple types of data from different sectors.[65] However, it is of the utmost importance that HHS seeks to improve its capacity to generate robust evidence, both to inform policymaking and to make the best use of government resources and investment.

The 2018 HHS Data Strategy is intended to address current gaps in the use of data for evidence-based policymaking through the six-part priority framework of (1) Improving Access to HHS Data, (2) Enhancing Administrative Data for Research, (3) Increasing Data Linkages Across Diverse Data Assets, (4) Modernizing Privacy Protections, (5) Increasing Data Policy Coordination and Information Sharing Across the Department, and (6) Building a 21st Century Data-Oriented Workforce. By following this strategic framework over the next eight years, HHS will be in a better position not only to routinely use and share data for research purposes, but also to participate in the growing efforts across the federal government to use data for policy research and evaluation.

IMPLEMENTATION

The 2018 HHS Data Strategy is intended to be a visionary document that guides the HHS Data Council agenda over the next six to eight years. The Council will develop an Action Plan

biennially that operationalizes the priorities, opportunities, and strategies prescribed in this guide. When developing and implementing action steps, the Council will take into account Department priorities and seek to prioritize actions that will advance these goals. In addition, the Council will take into account budgetary constraints, enterprise risk management, and existing OpDiv best practices and expertise to ensure that new investments are targeted, cost-effective, and have a meaningful impact. The Action Plans will ensure that the 2018 Data Strategy is flexible and responsive to changing administration priorities, Council priorities, and emerging areas in health and human services. The HHS Data Council will be developing the first Action Plan in the next few months and will implement it over 12 to 18 months from the date on which the 2018 HHS Data Strategy is officially released.

ACKNOWLEDGMENTS

ACF	Administration for Children and Families
ACYF	Administration on Children, Youth, and Families
ACL	Administration for Community Living
ADWG	HHS Administrative Data Working Group
AHRQ	Agency for Healthcare Research and Quality
ASA	Office of the Assistant Secretary for Administration
ASPA	Office of the Assistant Secretary for Public Affairs
ASPE	Office of the Assistant Secretary for Planning and Evaluation
CBHSQ	Center for Behavioral Health Statistics and Quality (SAMHSA)
CDC	Centers for Disease Control and Prevention
CEP	Commission on Evidence-Based Policymaking
CIPSEA	Confidential Information Protection and Statistical Efficiency Act of 2002
CMS	Centers for Medicare & Medicaid Services
CTO	Office of the Chief Technology Officer
DP	Data Policy Division (ASPE)
DUA	Data Use Agreement
EBPMA	Evidence-Based Policymaking Act of 2017
EDI	Enterprise Data Inventory
EDM	Enterprise Data Management
EHR	Electronic Health Record
ESS	European Statistical System
FCSM	Federal Committee on Statistical Methodology
FDA	Food and Drug Administration
FIPPS	Fair Information Practice Principles
FISMA	Federal Information Security Modernization Act of 2014
FITARA	Federal Information Technology Acquisition Reform Act
FSRDC	Federal Statistical Research Data Center

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
ICSP	Interagency Committee on Statistical Policy
IDS	Integrated Data Systems
IT	Information Technology
MOU	Memorandum of Understanding
NAS	National Academies of Sciences, Engineering, and Medicine
NCHS	National Center for Health Statistics
NHIS	National Health Interview Survey
NIDA	National Institute on Drug Abuse
NIH	National Institutes of Health
NIST	National Institute of Standards and Technology
NSDUH	National Survey on Drug Use and Health
NSDS	National Secure Data Service
NVSS	National Vital Statistics System
OCIO	Office of the Chief Information Officer
ONC	Office of the National Coordinator for IT
OMB	Office of Management and Budget
OPRE	Office of Planning, Research and Evaluation (ACF)
OpDiv	Operating Division (HHS)
PHI	Protected Health Information
PII	Personally Identifiable Information
PRA	Paperwork Reduction Act of 1995
PUF	Public Use File
RUF	Restricted Use File

RDC   Research Data Center

SAMHSA   Substance Abuse and Mental Health Administration

StaffDiv   Staffing Division (HHS)

APPENDIX A: 2011 HHS DATA STRATEGY PRINCIPLES

The value of HHS data depends on their relevance, timeliness, availability, and distinctive contributions. As HHS develops strategies and recommendations for data collection that move the Department forward in achieving these goals, the following principles will guide planning efforts.

Privacy, confidentiality, and security. HHS data describe people's behavior and experiences. We will ensure that all HHS data collection, analysis, and dissemination activities respect the privacy of individuals, protect the confidentiality of respondents' information, and minimize response burden. Some HHS data are collected about providers, health care institutions, plans, and grantees. In all cases, we will endeavor to carefully balance the need for transparency and confidentiality in an appropriate manner and within the confines of the law.

Relevance. We will continually review our data collection activities and adapt them to reflect changes in the range of HHS activities; developments in the sciences of medicine, medical practice, public health, and social services; and modifications in data collection activities within the Department, in other federal Departments, in states, and in the private sector.

Efficiency. We will evaluate, assess, and, where appropriate, employ new data collection strategies and technologies to ensure that our data are collected as expeditiously and cost-effectively as possible. We will be mindful of the need to minimize redundancy.

Availability and ease of use. HHS data have the potential to be useful to everyone. We will make the information we collect and our analyses of it available to many different types of users through multifaceted dissemination activities and improved user tools.

Innovation. Data collection and analysis benefit from advances in the sciences of data collection technologies, the growing availability of digital information, and statistical methods. We will contribute to and make full use of research that evaluates and improves methods, measures, and data collection programs.

Integration. In planning and conducting HHS data collection activities, we will promote opportunities for the coordination, integration, and alignment of HHS data collection activities in a manner that supports integration across systems; ensures quality, utility, and timeliness; and addresses high priority data gaps in a coordinated fashion. In particular, we will aim to align and coordinate data capabilities afforded by administrative data systems, EHR systems, and information exchange capabilities in meeting data needs and assuring the benefits of these initiatives.

Scientific integrity. As we seek improvements in timeliness and responsiveness to policy and program needs, we will be attentive to the requirements of quality, accuracy, and reliability.

APPENDIX B: U.S. PRIVACY POLICIES AND LAWS APPLICABLE TO HHS

What types of policies protect privacy?

The origins of modern data protection laws can be found in a 1973 advisory committee report to the Secretary of the U.S. Department of Health, Education, and Welfare. In its report, *Records, Computers, and the Rights of Citizens*, the Advisory Committee on Automated Personal Data Systems recommended adoption of a Code of Fair Information Practice identifying for the first time specified safeguards that, regardless of the type of data being managed, would protect individual privacy. These included individual rights, such as the right to obtain access, request amendment, and seek redress, and proposed obligations on federal agencies throughout the entire life cycle of information, from creation and collection, to use, sharing, and eventual disposal of data. The federal Privacy Act of 1974 was the first law to be based on this Code, and its structure was a model for future U.S. and international privacy protections.

In 2008, the Department of Homeland Security (DHS) articulated the original five elements of the Code as eight Fair Information Practice Principles (FIPPs): Transparency, Individual Participation, Purpose Specification, Data Minimization, Use Limitation, Data Quality and Integrity, Security, and Accountability and Auditing.[66] In so doing, DHS incorporated the term "personally identifiable information (PII)," that earlier had been defined by OMB in its guidance to agencies under the Electronic Government Act of 2002.[67] While the DHS articulation only applied to its own programs and systems, other agencies have found the formulation useful.

Many of the laws briefly described in Table B.1: U.S. Privacy Laws Applicable to HHS incorporate some or all of the FIPPs.

Together, the principles of Transparency (notice) and Individual Participation (the right to grant or withhold consent to collection) combine to produce what is often referred to as "informed consent." The use of informed consent respects individual decision making and autonomy by allowing individuals to know prior to revealing data about themselves how those data are to be used. For example, under the Federal Policy for the Protection of Human Subjects, better known as the "Common Rule," enrolling individuals in research requires legally effective informed consent obtained "under circumstances that provide the prospective subject … opportunity to consider whether or not to participate and that minimize the possibility of coercion or undue influence."

The data minimization principle is embodied in the Privacy Act of 1974, which directs each agency to "maintain in its records 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 by executive order of the President."[68]

Table B.1: U.S. Privacy Laws Applicable to HHS

Law/Document and Brief Description
The Privacy Act of 1974 (5 U.S.C. § 552a)
The Privacy Act of 1974 establishes a code of fair information practice that governs the collection, maintenance, use, and dissemination of information about individuals that is maintained by a federal agency in a "system of records." A system of records is a group of records about living, natural persons, under the control of an agency, from which information is retrieved by the name of the individual or by another identifier assigned to the individual. The Privacy Act requires U.S. government agencies to give public notice of the existence of systems of records by publication in the *Federal Register*. The Privacy Act prohibits disclosure from a system of records absent the written consent of the individual, unless the disclosure is pursuant to one of twelve statutory exceptions. Necessary exceptions may be added administratively by publication in the *Federal Register* and notice to the Congress. The Act also provides individuals with a means by which to access or request amendment of their records, to seek an accounting of disclosures, or to seek civil redress. It sets forth various agency record-keeping requirements, establishes criminal penalties for violation of these requirements, and governs computer matching when used to determine federal benefits or recouping payments. Agencies may choose to exempt themselves from some of these provisions for intelligence, law enforcement, or other necessary equities by rulemaking.
The Federal Information Security Modernization Act of 2014 (FISMA) (44 U.S.C. §§ 3551-3558)
FISMA requires each agency to develop, document, and implement an agency-wide information security program that includes plans and procedures for ensuring continuity of operations of information systems that support the operations and assets of the agency.
Electronic Government Act of 2002 (E-Gov, 44 U.S.C. § 3501 note)
The Electronic Government Act of 2002 recognized that advances in technology, networks, internet access, and communications also have important ramifications for the protection of personal information contained in government records and systems. Section 208 of E-Gov requires all federal agencies that develop or procure new IT, or that make substantial changes to existing IT, involving the collection, maintenance, or dissemination of "information in identifiable form,", to undergo a process of privacy impact assessment (PIA). A PIA is an analysis of how covered information is collected, stored, protected, shared, and managed. The purpose of a PIA is to demonstrate that system owners and developers have incorporated privacy protections throughout the entire life cycle of a system. The result of the analysis is memorialized in a document that is also known as a Privacy Impact Assessment. The Act requires an agency to make PIAs publicly available, except when an agency, in its discretion, determines that publication of the PIA would raise security concerns, reveal classified information, or compromise sensitive information (e.g., potentially damaging to the national interest, law enforcement efforts, or competitive business information contained in the assessment).

Law/Document and Brief Description
Confidential Information Protection and Statistical Efficiency Act (CIPSEA)   **(44 U.S.C Ch. 35, Subch. III)**
CIPSEA provides strong confidentiality protections for statistical information collections, such as surveys and censuses, as well as for other statistical activities, such as data analysis, modeling, and sample design, that are sponsored or conducted by federal agencies. It dictates that data or information acquired by an agency under a pledge of confidentiality and for exclusively statistical purposes shall be used exclusively for statistical purposes and only disclosed in identifiable form to another statistical agency for a statistical purpose, except with the informed consent of the subject. It establishes felony penalties for willfully disclosing information protected by CIPSEA to someone not entitled to receive it.
Paperwork Reduction Act of 1995 (PRA) (44 U.S.C. § Chapter 35 et seq)
The PRA, signed into law in 1980 and reauthorized in 1995, provides the statutory framework for the federal government's collection of information from the public. The goals of the PRA include (1) minimizing paperwork and reporting burdens on the American public and (2) ensuring the maximum possible utility from the information collected.   To support these goals, the PRA requires federal agencies to take specific steps before requiring or requesting information from the public, or requiring regulatory recordkeeping or disclosure. These steps include seeking public comment on proposed information collections and submitting proposed collections for review and approval by the Office of Management and Budget (OMB). Within OMB, the Office of Information and Regulatory Affairs (OIRA) carries out the information collection review.   Among the purposes of the PRA is to ensure that the creation, collection, maintenance, use, dissemination, and disposition of information by or for the federal government is consistent with applicable laws, including laws relating to privacy and confidentiality, such as the Privacy Act of 1974; security of information, and access to information, including under the Freedom of Information Act.
21st Century Cures Act (Public Law No: 114-255)
The 21st Century Cures Act is intended to accelerate the discovery, development, and delivery of 21st century cures, and for other purposes. It includes the strengthening of provisions governing the granting of "certificates of confidentiality" to research that collects individually identifiable data in order to protect that data from legal demands such as service of process, subpoenas, or court orders. The program mandates the Secretary of HHS to provide a Certificate of Confidentiality to any federally sponsored research, and permits the Secretary to provide a certificate voluntarily to non-federally funded research.
Federal Policy for the Protection of Human Subjects (Common Rule) (42 U.S.C. § 289)
The National Research Act of 1974 (Pub. L. 93-348) established the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research which, in 1979, published a very influential document, known as the Belmont Report, on which the current U.S. system of protection for human research subjects is based.   Congress enacted a law requiring that each entity that applies for a grant, contract, or cooperative agreement involving the conduct of biomedical or behavioral research involving human subjects set up an Institutional Review Board (IRB) to review the research in order to protect the rights of the human subjects.

Law/Document and Brief Description
Fifteen federal departments and agencies simultaneously promulgated coordinated regulations in 1991 regarding the protection of human subjects, which together became known as the: Common Rule." The HHS regulations, 45 CFR part 46, include four subparts: subpart A, also known as the Federal Policy or the Common Rule; subpart B, additional protections for pregnant women, human fetuses, and neonates; subpart C, additional protections for prisoners; and subpart D, additional protections for children. A fifth subpart, subpart E, which concerns registration of IRBs, was added in 2009. For all participating departments and agencies, the Common Rule outlines the basic provisions for IRBs, informed consent, and Assurances of Compliance. Human subject research is governed by the regulations of the department or agency supporting the research. The head of that department or agency retains final judgment as to whether a particular activity it conducts or supports is covered by the Common Rule. HHS and 15 other federal departments and agencies have issued final revisions to the Common Rule. The final Common Rule was published in the Federal Register on January 19, 2017. It implements new steps to better protect human subjects involved in research while facilitating valuable research and reducing burden, delay, and ambiguity for investigators.
Health Insurance Portability and Accountability Act, as amended (HIPAA Privacy, Security, and Breach Notification Rules) (45 C.F.R. Parts 160 and 164)
HIPAA Privacy Rule The HIPAA Privacy Rule, 45 C.F.R. Part 164, Subpart E, adopted by HHS pursuant to HIPAA, establishes national standards to protect individuals' health information, called "protected health information" (PHI). HIPAA applies to covered entities, including health plans, health care clearinghouses, and those health care providers that conduct certain health care transactions electronically as well as their business associates. Within HHS, there are four covered entities: the Centers for Medicare and Medicaid Services, Indian Health Service, HHS components administering health benefits for the Commissioned Corps, and The World Trade Center Health Program. A major purpose of the Privacy Rule is to define and limit the circumstances in which an individual's PHI may be used or disclosed by covered entities. A covered entity may not use or disclose protected health information, except as the Privacy Rule permits or requires or as the individual who is the subject of the information authorizes in writing. The Privacy Rule provides patients with certain rights, including the right to examine and obtain a copy of their health records and to request corrections. A central aspect of the Privacy Rule is the principle of "minimum necessary" use and disclosure. A covered entity must make reasonable efforts to use, disclose, and request only the minimum amount of PHI needed to accomplish the intended purpose of the use, disclosure, or request. *HIPAA Security Rule* The HIPAA Security Rule, 45 C.F.R. Part 164, subpart C, adopted by HHS pursuant to HIPAA, establishes national standards to protect individuals' electronic PHI that is created, received, used, or maintained by a covered entity or business associate. The Security Rule requires appropriate administrative, physical, and technical safeguards to ensure the confidentiality, integrity, and security of electronic PHI.

Law/Document and Brief Description

HIPAA Breach Notification Rule

The HIPAA Breach Notification Rule, 45 C.F.R. Part 164, subpart D, requires HIPAA-covered entities and their business associates to provide notification following a breach of unsecured PHI.

Drug Abuse Prevention, Treatment, and Rehabilitation Act (Confidentiality of Alcohol and Drug Abuse Patient Records) (42 U.S.C § 290dd–2 and 42. C.F.R. Part 2)

Confidentiality of patient records of the identity, diagnosis, prognosis, or treatment of a substance use disorder (alcohol and drug abuse) is governed by 42 U.S.C § 290dd–2 and 42 C.F.R Part 2. The statute and regulation require that records related to patients' substance use disorders remain confidential and subject to certain specific exceptions or to patients' consent for disclosing such information. The statute extends to cover "any program or activity relating to substance abuse education, prevention, training, treatment, rehabilitation, or research, which is conducted, regulated, or directly or indirectly assisted by any department or agency of the United States."

Title 13 U. S. Code (Census Bureau)

The U.S. Census Bureau is bound by Title 13 of the United States Code. These laws not only provide authority for the work of the Census Bureau, but provide strong protection for the information the Census Bureau collects from individuals and businesses: Private information is never published. It is against the law to disclose or publish any private information that identifies an individual or business, including names, addresses (including GPS coordinates), Social Security numbers, and telephone numbers.

- The Census Bureau collects information to produce statistics, and the information may only be disclosed in de-identified, aggregate form for the purpose of statistical activities or reporting.

- Personal information cannot be used against respondents by any government agency or court.

- Census Bureau employees are sworn to protect confidentiality. People sworn to uphold Title 13 are legally required to maintain the confidentiality of Census data. Every person with access to Census data is sworn for life to protect the information they had access to during their tenure at the Census Bureau.

- Violating the law is a serious federal crime. Anyone who violates this law will face severe penalties, including a federal prison sentence of up to five years, a fine of up to $250,000, or both.

Reference: The Census Bureau,
https://www.census.gov/history/www/reference/privacy_confidentiality/title_13_us_code.html

Federal Information Technology Acquisition Reform Act (FITARA)

FITARA was enacted on December 19, 2014, and outlines specific requirements related to (1) CIO authority enhancements; (2) enhanced transparency and improved risk management in IT investments; (3) portfolio reviews; (4) expansion of training and use of IT cadres; (5) the Federal Data Center Consolidation Initiative (FDCCI), more recently known as the Data Center Optimization Initiative (DCOI); (6) maximizing the benefit of the federal strategic sourcing initiative (SSI); and (7) a government-wide software purchasing program. The goal of FITARA is to eliminate duplication and waste in the acquisition of IT for the federal government. The goals of the legislation are reducing duplicative systems, examining software licensing options, making the business case for acquisition, and the consolidation of data centers. The objective of FITARA is to improve the management of IT within an agency and, hence, improve the ability for that agency to deliver its mission and conduct its business.

Reference: TTS Handbook, https://handbook.18f.gov/fitara/

Law/Document and Brief Description
National Center for Health Statistics (NCHS) Confidentiality Statute (42 U.S.C. 242m(d))
NCHS is designated as a Federal Statistical Unit and complies with the November 2014 OMB Statistical Policy Directive No. 1, The legislation that created NCHS and that authorizes data collection, the Public Health Service Act, contains a provision (42 U.S.C. § 242m(d)) that that prohibits NCHS from using any personal information collected by survey for any purpose other than what was described to survey participants and from sharing that information with anyone not clearly mentioned to the participants. This provision enables NCHS to ensure that their survey respondents' information remains strictly confidential. Reference: National Center for Health Statistics, https://www.cdc.gov/nchs/about/policy/confidentiality.htm
SAMHSA, Center for Behavioral Health Statistics and Quality (CBHSQ) **(42 U.S.C. § 290aa(p))**
CBHSQ is designated as a Federal Statistical Unit and complies with the November 2014 OMB Statistical Policy Directive No. 1, and its data collection authority/mandate is from Section 505 of the Public Health Service Act (42 U.S.C. § 290aa-4). That provision limits the use of identifiable information collected by SAMHSA for statistical purposes to the production of statistical information, and prohibits disclosure or publication of identifiable information, unless with the consent of the subject. Reference: SAMHSA's Center for Behavioral Health Statistics and Quality, https://www.samhsa.gov/data/about-us Cornell Law Library, https://www.law.cornell.edu/uscode/text/42/290aa
Cybersecurity Information Sharing Act of 2015 (CISA) **(6 U.S.C. §§ 149, 151, 1501-1510, 1521-1525, 1531-1533)**
Congress enacted CISA, Title I of the Cybersecurity Act, to direct the Department of Homeland Security, in collaboration with other named agencies, including HHS, to create a voluntary cybersecurity information sharing process that will protect participants from certain types of liability and encourage public and private entities to share cyber threat information in real time while protecting the privacy and civil liberties of individuals.
The Children's Online Privacy Protection Act (COPPA) (15 U.S.C. §§ 6501-6505)
COPPA imposes certain requirements on operators of websites or online services directed to children under 13 years of age, and on operators of other websites or online services that have actual knowledge that they are collecting personal information online from a child under 13 years of age. Federal agencies, including HHS, are also subject to this law. COPPA requires operators of websites to ask visitors to their sites whether they are over the age of 13, and, if not, direct them to obtain the consent of a parent or other responsible adult prior to obtaining access.

Law/Document and Brief Description
OMB Circular A-108 Federal Agency Responsibilities for Review, Reporting, and Publication under the Privacy Act
This OMB Circular, re-released in December 2016, Provides guidance on implementation and reporting requirements under the Privacy Act of 1974. It covers reporting requirements on new and revised systems of records, new or revised routine uses, computer matching programs, the documents required for submission to OMB and the Congress when requesting review, and the timing and deadlines associated with such requests. Reference: OMB, https://www.whitehouse.gov/sites/whitehouse.gov/files/omb/circulars/A108/omb_circular_a-108.pdf
OMB M-07-16 Safeguarding Against and Responding to the Breach of Personally Identifiable Information.
This Memorandum, issued by OMB in May 2007, sets out a framework intended to reduce risks related to a data breach of personally identifiable information. It requires agencies to develop and publish a breach notification policy. The policy must include requirements for reporting and handling incidents, external breach notification, and the responsibilities of individuals authorized to access personally identifiable information. Reference: OMB, https://www.whitehouse.gov/sites/whitehouse.gov/files/omb/memoranda/2007/m07-16.pdf
The National Institute of Standards and Technology (NIST)
NIST Special Publication: Security and Privacy Controls for Federal Information Systems and Organizations This publication provides a catalog of security and privacy controls for federal information systems and organizations and a process for selecting controls to protect organizational operations (including mission, functions, image, and reputation), organizational assets, individuals, other organizations, and the Nation from a diverse set of threats including hostile cyber attacks, natural disasters, structural failures, and human errors (both intentional and unintentional). The security and privacy controls are customizable and implemented as part of an organization-wide process that manages information security and privacy risk. The controls address a diverse set of security and privacy requirements across the federal government and critical infrastructure, derived from legislation, Executive Orders, policies, directives, regulations, standards, and/or mission/business needs. The publication also describes how to develop specialized sets of controls, or overlays, tailored for specific types of missions/business functions, technologies, or environments of operation. Finally, the catalog of security controls addresses security from both a functionality perspective (the strength of security functions and mechanisms provided) and an assurance perspective (the measures of confidence in the implemented security capability). Addressing both security functionality and assurance helps to ensure that information technology component products and the information systems built from those products using sound system and security engineering principles are sufficiently trustworthy.

Law/Document and Brief Description
Reference: NIST, https://csrc.nist.gov/publications/detail/sp/800-53/rev-4/final
NIST Cybersecurity Framework   This voluntary Framework consists of standards, guidelines, and best practices for managing cybersecurity-related risk. The Framework's prioritized, flexible, and cost-effective approach helps to promote the protection and resilience of critical infrastructure and other sectors important to the economy and national security.   Reference: NIST, https://www.nist.gov/cyberframework

BIBLIOGRAPHY

Agency for Healthcare Research and Quality. (2018). *Healthcare cost and utilization project.* Retrieved on August 29, 2018, from https://hcupnet.ahrq.gov.

Ahmad F.B., Rossen, L.M., Spencer, M.R., Warner, M., & P. Sutton. (2018). *Provisional drug overdose death counts.* Atlanta, GA: National Center for Health Statistics. Retrieved on August 29, 2018, from https://www.cdc.gov/nchs/nvss/vsrr/drug-overdose-data.htm.

Centers for Disease Control and Prevention (2018). Access modes. Retrieved on August 29, 2018, from https://www.cdc.gov/rdc/b2accessmod/acs200.htm.

Centers for Disease Control and Prevention. (2018). CDC Wonder. Retrieved on August 29, 2018, from https://wonder.cdc.gov/.

Centers for Disease Control and Prevention. National Center for Health Statistics. (2018). *The National Health Interview Survey Early Release Program.* Retrieved on August 29, 2018, from https://www.cdc.gov/nchs/nhis/releases.htm.

Centers for Disease Control and Prevention. (2017). *Improving disease surveillance at CDC—a strategic approach.* Retrieved on August 29, 2018, from https://www.cdc.gov/surveillance/Improving-Surveillance-Background.html.

Centers for Disease Control and Prevention, National Center for Health Statistics. (2017, November 9). *NCHS data linkage.* Retrieved on August 29, 2018, from https://www.cdc.gov/nchs/data-linkage/index.htm.

Centers for Disease Control and Prevention (2015, December 15). NCHS Research Data Center. Retrieved on August 29, 2018, from https://www.cdc.gov/rdc/index.htm.

The Commission on Evidence-Based Policymaking. (2017, September). *The promise of evidence-based policymaking: Report of the Commission on Evidence-Based Policymaking* (pp. 101-102). Retrieved on August 29, 2018, from https://www.cep.gov/.

Czajka, J., Schneider, C., Sukasih, A., & Collins, K. (2014, September). *Minimizing disclosure risk in HHS open data initiatives.* Washington, DC: Mathematica Policy Research, September 2014. Retrieved on August 29, 2018, from https://aspe.hhs.gov/system/files/pdf/77196/rpt_Disclosure.pdf.

Dahn, Shannon. Fair information practice principles [PowerPoint slides]. Presentation by the Department of Homeland Security. Retrieved on August 29, 2018, from https://www.dhs.gov/publication/fair-information-practice-principles-fipps.

Data Sharing for Demographic Research. (2018). *What is the virtual data enclave (VDE)?* Retrieved on August 29, 2018, from https://www.icpsr.umich.edu/icpsrweb/content/DSDR/what-is-the-vde.html.

Department of Health and Human Services. (2018). *Agency priority goal action plan: Combined data analyses.* Retrieved on August 29, 2018, from https://www.performance.gov/health_and_human_services/APG_hhs_3.html.

Department of Health and Human Services. (2018). HHS agencies & offices. Retrieved on September 6, 2018, from https://www.hhs.gov/about/agencies/hhs-agencies-and-offices/index.html.

Department of Health and Human Services. (2018). Message from the Secretary, 2018–2022 HHS Strategic Plan. Retrieved on August 29, 2018, from https://www.hhs.gov/about/strategic-plan/message/index.html.

Department of Health and Human Services. (2018). Strategic Plan FY 2018 – 2022. Washington, DC: Department of Health and Human Services. Retrieved on August 29, 2018, from https://www.hhs.gov/about/strategic-plan/index.html.

Department of Health and Human Services (2018, August 16). HHS Enterprise Data Inventory. Retrieved on August 29, 2018, from https://healthdata.gov/dataset/hhs-enterprise-data-inventory.

Department of Health and Human Services. (2018, February 28). Strategic Plan FY 2018 – 2022. Strategic Objective 5.3. Optimize information technology investments to improve process efficiency and enable innovation to advance program mission goals. Retrieved on August 29, 2018, from https://www.hhs.gov/about/strategic-plan/strategic-goal-5/index.html.

Department of Health and Human Services. (2018, February 12). *HHS Secretary Azar Statement on President Trump's FY 2019 Budget.* Press release. Retrieved on August 29, 2019, from https://www.hhs.gov/about/news/2018/02/12/hhs-secretary-azar-statement-president-trumps-fy-2019-budget.html.

Department of Health and Human Services. (2015). Guidance regarding methods for de-identification of protected health information in accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. Retrieved on August 29, 2018, from https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html.

Department of Health and Human Services. (2015). Guiding principles and common approach for enhancing public access to the results of research funded by HHS operating divisions. Retrieved on August 29, 2018, from https://www.hhs.gov/open/public-access-guiding-principles/index.html.

Department of Health and Human Services. Division of Data Policy. (n.d.). Retrieved on August 29, 2018, from https://aspe.hhs.gov/division-data-policy.

Department of Health and Human Services. Office of the Chief Information Officer. (2017, September 8). Retrieved on August 29, 2018, from https://www.hhs.gov/about/agencies/asa/ocio/index.html.

Department of Health and Human Services. Office of the Chief Technology Officer. (2017, May 17). Retrieved on August 29, 2018, from https://www.hhs.gov/about/agencies/cto/index.html.

Department of Health and Human Services, Office of the Inspector General. (2016, June 7). *What role does data play in fighting healthcare fraud, waste and abuse?* Podcast transcript retrieved on August 29, 2018, from https://oig.hhs.gov/newsroom/podcasts/2016/cdo-trans.asp.

Department of Health and Human Services. Office of the National Coordinator for Health Information Technology. (2018, May 22). Retrieved on August 29, 2018, from https://www.healthit.gov/topic/about-onc.

Eurostat. (2015). *ESS handbook for quality reports.* 2014 edition. Luxembourg: Publications Office of the European Union. Retrieved on August 29, 2018, from http://ec.europa.eu/eurostat/documents/3859598/6651706/KS-GQ-15-003-EN-N.pdf.

Floridi, L., & Taddeo, M. What is data ethics? *Philosophical Transactions of the Royal Society A, Mathematical, Physical, and Engineering Sciences, 374*(2083). doi:10.1098/rsta.2016.0360.

Iwig, W., Berning, M., Marck, P., & Prell, M. (2013). Federal Committee on Statistical Methodology (FCSM). *Data quality assessment tool for administrative data.* Retrieved on August 29, 2018, from https://nces.ed.gov/FCSM/pdf/DataQualityAssessmentTool.pdf.

Maxwell, K. (2017, February). *Issues in accessing and using administrative data.* OPRE Report #2017-24. Washington, DC: Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research, and Evaluation. Retrieved on August 29, 2018, from https://www.acf.hhs.gov/sites/default/files/opre/issues_in_accessing_and_using_administrative_data_3617_508.pdf.

Millar, A., Czajka, J., Morris, R., Tallapragada, K., Lauffer, S., & Fisher, S. (2017, September). *Administrative data assessment environmental scan report*. Washington, DC: Mathematica Policy Research.

National Academies of Sciences, Engineering, and Medicine. (2018). Knowledge for data scientists (p. 9). In *Data science for undergraduates: opportunities and options.* Washington, DC: The National Academies Press. doi.org/10.17226/25104.

National Academies of Sciences, Engineering, and Medicine. (2017). *Federal statistics, multiple data sources, and privacy protection: Next steps*. Washington, DC: The National Academies Press. doi.org/10.17226/24893.

National Academies of Sciences, Engineering, and Medicine. (2017). *Innovations in federal statistics: Combining data sources while protecting privacy and federal statistics, multiple data sources, and privacy protection: Next steps.* Washington, DC: The National Academies Press. doi: 10.17226/24652.

National Academies of Science, Engineering, Medicine. (2017). *Principles and practices for a federal statistical agency* (6th ed.). Retrieved on August 29, 2018, from https://www.nap.edu/catalog/24810/principles-and-practices-for-a-federal-statistical-agency-sixth-edition.

National Institute of Health. (2018). *Reimagine HHS: Moving towards a 21st century workforce.* Retrieved on August 29, 2018, from https://hr.nih.gov/about/events/hhs-hr-clinic/presentations/reimagine-hhs-moving-towards-21st-century-workforce.

National Library of Medicine. (2018, March 17). HSRR (Health Services and Sciences Research Resources). Retrieved on August 29, 2018, from https://hsrr.nlm.nih.gov.

Office of Management and Budget (2013, May 9). Memorandum M-13-13: Open Data Policy-Managing Information as an Asset. Project Open Data. Retrieved on September 13, 2018, from https://www.whitehouse.gov/sites/whitehouse.gov/files/omb/memoranda/2013/m-13-13.pdf.

Office of Management and Budget. (2016, July 16). Memorandum to the Heads of Executive Departments and Agencies. *Management's responsibility for enterprise risk management and internal control.* OMB Circular No. A-23. Retrieved on August 29, 2018, from https://www.whitehouse.gov/sites/whitehouse.gov/files/omb/memoranda/2016/m-16-17.pdf.

Office of Management and Budget. (2016, July 15). *Barriers to using administrative data for evidence-building.* White paper prepared by OMB staff for the Commission on Evidence-Based Policymaking. Washington, DC: Office of Management and Budget.

Office of Management and Budget (Memorandum M-14-06). (2014). *Guidance for providing and using administrative data for statistical purposes.* Washington, DC: Office of Management and Budget.

Office of Management and Budget. (2014, December 2). Statistical policy directive no. 1: Fundamental responsibilities of federal statistical agencies and recognized statistical units; Notice. *Federal Register, 79*(231). Retrieved on August 29, 2018, from https://www.gpo.gov/fdsys/pkg/FR-2014-12-02/pdf/2014-28326.pdf.

President's Management Council and the Executive Office of the President. (2018). *President's management agenda: Modernizing government for the 21st century.* Retrieved on June 12, 2018, from https://www.whitehouse.gov/wp-content/uploads/2018/03/Presidents-Management-Agenda.pdf.

Rothwell, C. (2018, August). NCHS director addresses challenges for federal statistical agencies. *AMSTATNEWS.* Retrieved on August 29, 2018, from http://magazine.amstat.org/blog/2018/08/01/nchsdirector/.

Seeskin, Z.H., LeClere, F.B., Ahn, J., & Williams, J.A. (2018, July 30). Uses of alternative data sources for public health statistics and health policymaking: challenges and opportunities. *Joint Statistical Meetings.* Vancouver, BC.

Siddiqui, Mona, and Rashida Dorsey. (n.d.). *Agency priority goal action plan: Combined data analyses.* Retrieved on August 29, 2018, from https://www.performance.gov/health_and_human_services/FY2018_Q1_HHS_Combined_Data_Analyses.pdf.

Smart, R., Kase, C.A., Meyer, A., & Stein, B.D. (2018). *Research to address the opioid crisis: Approaches to data linkage.* Submitted to: Office of Health Policy, Assistant Secretary for Planning and Evaluation, U.S. Department of Health and Human Services. Santa Monica, CA: RAND Corporation. Publication expected in September 2018.

Wo, L. Chang, Roy, A., Grady, N., Reinsch, R., Underwood, M., Fox, G., Boyd, D., & von Laszewski, G. (2018, June 26). *NIST big data interoperability framework: Volume 1, definitions (version 2).* Special Publication (NIST SP)-1500-1-9. Retrieved on August 29, 2018, from https://www.nist.gov/publications/nist-big-data-interoperability-framework.

Zook, M., Barocas, S., Boyd, D., Crawford, K., Keller, E., Gangadharan, S.P., . . . & Frank Pasquale. (2017). Ten simple rules for responsible big data research. *PLoS Computational Biology, 13*(3). doi:10.1371/journal.pcbi.1005399.

ENDNOTES

[1] President's Management Council and the Executive Office of the President. (2018). *President's management agenda: Modernizing government for the 21st century.* Retrieved on June 12, 2018, from https://www.whitehouse.gov/wp-content/uploads/2018/03/Presidents-Management-Agenda.pdf.

[2] National Institute of Health. (2018). *ReImagine HHS: Moving towards a 21st century workforce.* Retrieved on August 29, 2018, from https://hr.nih.gov/about/events/hhs-hr-clinic/presentations/reimagine-hhs-moving-towards-21st-century-workforce.

[3] Department of Health and Human Services. (2018). *Agency priority goal action plan: Combined data analyses.* Retrieved on August 29, 2018, from https://www.performance.gov/health_and_human_services/APG_hhs_3.html.

[4] The Commission on Evidence-Based Policymaking. (2017, September). *The promise of evidence-based policymaking: Report of the commission on evidence-based policymaking.* Retrieved on August 29, 2018, from https://www.cep.gov/.

[5] The Commission on Evidence-Based Policymaking. (2017, September). *The promise of evidence-based policymaking: Report of the commission on evidence-based policymaking.* Retrieved on August 29, 2018, from https://www.cep.gov/.

[6] President's Management Council and the Executive Office of the President. (2018). *President's management agenda: Modernizing government for the 21st century.* Retrieved on June 12, 2018, from https://www.whitehouse.gov/wp-content/uploads/2018/03/Presidents-Management-Agenda.pdf.

[7] Department of Health and Human Services. (2018). Message from the Secretary, 2018–2022 HHS Strategic Plan. Retrieved on August 29, 2018, from https://www.hhs.gov/about/strategic-plan/message/index.html.

[8] Office of Management and Budget (Memorandum M-14-06). (2014). *Guidance for providing and using administrative data for statistical purposes.* Washington, DC: Office of Management and Budget.

[9] Department of Health and Human Services. (2018, February 12). *HHS Secretary Azar Statement on President Trump's FY 2019 Budget.* Press release. Retrieved on August 29, 2019, from https://www.hhs.gov/about/news/2018/02/12/hhs-secretary-azar-statement-president-trumps-fy-2019-budget.html.

[10] Department of Health and Human Services. (2018). Strategic Plan FY 2018 – 2022. Washington, DC: Department of Health and Human Services. Retrieved on August 29, 2018, from https://www.hhs.gov/about/strategic-plan/index.html.

[11] Maxwell, K. (2017, February). *Issues in accessing and using administrative data.* OPRE Report #2017-24. Washington, DC: Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research, and Evaluation. Retrieved on August 29, 2018, from https://www.acf.hhs.gov/sites/default/files/opre/issues_in_accessing_and_using_administrative_data_3617_508.pdf.

[12] Centers for Disease Control and Prevention. (2018). CDC Wonder. Retrieved on August 29, 2018, from https://wonder.cdc.gov/.

[13] Agency for Healthcare Research and Quality. (2018). *Healthcare cost and utilization project.* Retrieved on August 29, 2018, from https://hcupnet.ahrq.gov.

[14] The term "PII," as defined in OMB Circular A-130, refers to "information that can be used to distinguish or trace an individual's identity, either alone or when combined with other information that is linked or linkable to a specific individual." Definition from: https://www.gsa.gov/reference/gsa-privacy-program/rules-and-policies-protecting-pii-privacy-act.

[15] As defined in the HIPAA Privacy Rule, PHI refers to "individually identifiable health information" held or transmitted by a covered entity or its business associate, in any form or medium, whether electronic, on paper, or oral. Definition from: https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html#protected.

[16] Data Sharing for Demographic Research. (2018). *What is the virtual data enclave (VDE)?* Retrieved on August 29, 2018, from https://www.icpsr.umich.edu/icpsrweb/content/DSDR/what-is-the-vde.html.

[17] Centers for Disease Control and Prevention (2015, December 15). NCHS Research Data Center. Retrieved on August 29, 2018, from https://www.cdc.gov/rdc/index.htm.

[18] Centers for Disease Control and Prevention (2018). Access modes. Retrieved on August 29, 2018, from https://www.cdc.gov/rdc/b2accessmod/acs200.htm.

[19] Department of Health and Human Services. (2015). *Guiding principles and common approach for enhancing public access to the results of research funded by HHS operating divisions.* Retrieved on August 29, 2018, from https://www.hhs.gov/open/public-access-guiding-principles/index.html.

[20] National Library of Medicine. (2018, March 17). HSRR (Health Services and Sciences Research Resources). Retrieved on August 29, 2018, from https://hsrr.nlm.nih.gov.

[21] Department of Health and Human Services. (2018, February 28). Strategic Plan FY 2018 – 2022. Strategic Objective 5.3. Optimize information technology investments to improve process efficiency and enable innovation to advance program mission goals. Retrieved on August 29, 2018, from https://www.hhs.gov/about/strategic-plan/strategic-goal-5/index.html.

[22] Office of Management and Budget (2013, May 9). Memorandum M-13-13: Open Data Policy-Managing Information as an Asset. Project Open Data. Retrieved on September 13, 2018, from https://www.whitehouse.gov/sites/whitehouse.gov/files/omb/memoranda/2013/m-13-13.pdf.

[23] Department of Health and Human Services (2018, August 16). HHS Enterprise Data Inventory. Retrieved on August 29, 2018, from https://healthdata.gov/dataset/hhs-enterprise-data-inventory

[24] Millar, A., Czajka, J., Morris, R., Kalyani Tallapragada, K., Lauffer, S., & Fisher, S. (2017, September). *Administrative data assessment environmental scan report*. Washington, DC: Mathematica Policy Research.

[25] Office of Management and Budget (2014). *Guidance for providing and using administrative data for statistical purposes*. OMB M-14-06. Washington, DC: Office of Management and Budget.

[26] Office of Management and Budget. (2016, July 15). *Barriers to using administrative data for evidence-building*. White paper prepared by OMB staff for the Commission on Evidence-Based Policymaking. Washington, DC: Office of Management and Budget.

[27] Czajka, J., Schneider, C., Sukasih, A., & Collins, K. (2014, September). *Minimizing disclosure risk in HHS open data initiatives.* Washington, DC: Mathematica Policy Research, September 2014. Retrieved on August 29, 2018, from https://aspe.hhs.gov/system/files/pdf/77196/rpt_Disclosure.pdf.

[28] Office of Management and Budget. (2014). *Guidance for providing and using administrative data for statistical purposes.* OMB M-14-06. Washington, DC: Office of Management and Budget.

[29] Millar, A., Czajka, J., Morris, R., Tallapragada, K., Lauffer, S., &Fisher, S. (2017, September). *Administrative data assessment environmental scan report*. Washington, DC: Mathematica Policy Research.

[30] Millar, A., Czajka, J., Morris, R., Tallapragada, K., Lauffer, S., & Fisher, S. (2017, September). *Administrative data assessment environmental scan report*. Washington, DC: Mathematica Policy Research

[31] Millar, A., Czajka, J., Morris, R., Tallapragada, K., Lauffer, S., & Fisher, S. *Administrative data assessment environmental scan report.* Washington, DC: Mathematica Policy Research, September 2017.

[32] Millar, A., Czajka, J., Morris, R., Tallapragada, K., Lauffer, S., & Fisher, S. *Administrative data assessment environmental scan report.* Washington, DC: Mathematica Policy Research, September 2017.

[33] Iwig, W., Berning, M., Marck, P., and Prell, M. (2013). Federal Committee on Statistical Methodology (FCSM). *Data quality assessment tool for administrative data.* Retrieved on August 29, 2018, from https://nces.ed.gov/FCSM/pdf/DataQualityAssessmentTool.pdf.

[34] Eurostat. (2015). *ESS handbook for quality reports.* 2014 edition. Luxembourg: Publications Office of the European Union. Retrieved on August 29, 2018, from http://ec.europa.eu/eurostat/documents/3859598/6651706/KS-GQ-15-003-EN-N.pdf.

[35] National Academies of Sciences, Engineering, and Medicine. (2017). *Federal statistics, multiple data sources, and privacy protection: Next steps*. Washington, DC: The National Academies Press. doi.org/10.17226/24893.

[36] Smart, R., Kase, C.A., Meyer, A., & Stein, B.D. (2018). *Research to address the opioid crisis: Approaches to data linkage.* Submitted to: Office of Health Policy, Assistant Secretary for Planning and Evaluation, U.S. Department of Health and Human Services. Santa Monica, CA: RAND Corporation. Publication expected in September 2018.

[37] National Academies of Sciences, Engineering, and Medicine. (2017). *Federal statistics, multiple data sources, and privacy protection: next steps*. Washington, DC: The National Academies Press. doi.org/10.17226/24893.

[38] Smart, R., Kase, C.A., Meyer, A., & Stein, B.D. (2018). *Research to address the opioid crisis: Approaches to data linkage.* Submitted to: Office of Health Policy, Assistant Secretary for Planning and Evaluation, U.S. Department of Health and Human Services. Santa Monica, CA: RAND Corporation. Publication expected in September 2018.

[39] Smart, R., Kase, C.A., Meyer, A., & Stein, B.D. (2018). *Research to address the opioid crisis: Approaches to data linkage.* Submitted to: Office of Health Policy, Assistant Secretary for Planning and Evaluation, U.S. Department of Health and Human Services. Santa Monica, CA: RAND Corporation. Publication expected in September 2018.

[40] Centers for Disease Control and Prevention, National Center for Health Statistics. (2017, November 9). *NCHS data linkage.* Retrieved on August 29, 2018, from https://www.cdc.gov/nchs/data-linkage/index.htm.

[41] Czajka, J., Schneider, C, Sukasih, A., Collins, K. (2014, September). *Minimizing disclosure risk in HHS open data initiatives.* Washington, DC: Mathematica Policy Research.

[42] Siddiqui, Mona, and Rashida Dorsey. (2018). *Agency priority goal action plan: Combined data analyses.* Retrieved on August 29, 2018, from https://www.performance.gov/health_and_human_services/FY2018_Q1_HHS_Combined_Data_Analyses.pdf.

[43] Office of Management and Budget. (2014, December 2). Statistical policy directive no. 1: Fundamental responsibilities of federal statistical agencies and recognized statistical units; Notice. *Federal Register, 79*(231). Retrieved on August 29, 2018, from https://www.gpo.gov/fdsys/pkg/FR-2014-12-02/pdf/2014-28326.pdf.

[44] Department of Health and Human Services. HHS agencies & offices. (2018). HHS.gov. Retrieved on September 6, 2018, from https://www.hhs.gov/about/agencies/hhs-agencies-and-offices/index.html.

[45] Office of Management and Budget. (2016, July 16). Memorandum to the Heads of Executive Departments and Agencies. *Management's responsibility for enterprise risk management and internal control.* OMB Circular No. A-23. Retrieved on August 29, 2018, from https://www.whitehouse.gov/sites/whitehouse.gov/files/omb/memoranda/2016/m-16-17.pdf.

[46] National Academies of Science, Engineering, Medicine. *Principles and practices for a federal statistical agency* (6th ed.). Retrieved on August 29, 2018, from https://www.nap.edu/catalog/24810/principles-and-practices-for-a-federal-statistical-agency-sixth-edition.

[47] Department of Health and Human Services. (2018). *Agency priority goal action plan: Combined data analyses.* Retrieved on August 29, 2018, from https://www.performance.gov/health_and_human_services/APG_hhs_3.html.

[48] Department of Health and Human Services. Office of the Chief Information Officer. (2017, September 8). Retrieved on August 29, 2018, from https://www.hhs.gov/about/agencies/asa/ocio/index.html.

[49] Department of Health and Human Services. Division of Data Policy. (n.d.). Retrieved on August 29, 2018, from https://aspe.hhs.gov/division-data-policy.

[50] Department of Health and Human Services. Office of the National Coordinator for Health Information Technology. (2018, May 22). Retrieved on August 29, 2018, from https://www.healthit.gov/topic/about-onc

[51] Department of Health and Human Services. Office of the Chief Technology Officer. (2017, May 17). Retrieved on August 29, 2018, from https://www.hhs.gov/about/agencies/cto/index.html.

[52] Centers for Disease Control and Prevention. National Center for Health Statistics. (2018). *The National Health Interview Survey Early Release Program.* Retrieved on August 29, 2018, from https://www.cdc.gov/nchs/nhis/releases.htm.

[53] Ahmad F.B., Rossen, L.M., Spencer, M.R., Warner, M., & P. Sutton. (2018). *Provisional drug overdose death counts.* Atlanta, GA: National Center for Health Statistics. Retrieved on August 29, 2018, from https://www.cdc.gov/nchs/nvss/vsrr/drug-overdose-data.htm.

[54] National Academies of Sciences, Engineering, and Medicine. (2018). Knowledge for data scientists (p. 9). In *Data science for undergraduates: opportunities and options.* Washington, DC: The National Academies Press. doi.org/10.17226/25104.

[55] President's Management Council and the Executive Office of the President. (2018). *President's management agenda: Modernizing government for the 21st century.* Retrieved on June 12, 2018, from https://www.whitehouse.gov/wp-content/uploads/2018/03/Presidents-Management-Agenda.pdf.

[56] The Commission on Evidence-Based Policymaking. (2017, September). *The promise of evidence-based policymaking: Report of the commission on evidence-based policymaking* (pp. 101-102). Retrieved on August 29, 2018, from https://www.cep.gov/.

[57] As noted in the NIH Strategic Plan for Data Science, data science is the interdisciplinary field of inquiry in which quantitative and analytical approaches, processes, and systems are developed and used to extract knowledge and insights from increasingly large and/or complex sets of data. (The NIH Strategic Plan for Data Science can be found at https://datascience.nih.gov/sites/default/files/NIH_Strategic_Plan_for_Data_Science_Final_508.pdf).

[58] Floridi, L., & Taddeo, M. What is data ethics? *Philosophical Transactions of the Royal Society A, Mathematical, Physical, and Engineering Sciences, 374*(2083). doi:10.1098/rsta.2016.0360.

[59] Zook, M., Barocas, S., Boyd, D., Crawford, K., Keller, E., Gangadharan, S.P., . . . & Frank Pasquale. (2017). Ten simple rules for responsible big data research. *PLoS Computational Biology, 13*(3). doi:10.1371/journal.pcbi.1005399.

[60] Department of Health and Human Services, Office of the Inspector General. (2016, June 7). *What role does data play in fighting healthcare fraud, waste and abuse?* Podcast transcript retrieved on August 29, 2018, from https://oig.hhs.gov/newsroom/podcasts/2016/cdo-trans.asp.

[61] Wo, L. Chang, Roy, A., Grady, N., Reinsch, R., Underwood, M., Fox, G., Boyd, D., & von Laszewski, G. (2018, June 26). *NIST big data interoperability framework: Volume 1, definitions (version 2)*. Special Publication (NIST SP)-1500-1-9. Retrieved on August 29, 2018, from https://www.nist.gov/publications/nist-big-data-interoperability-framework.

[62] Seeskin, Z.H., LeClere, F.B., Ahn, J., & Williams, J.A. (2018, July 30). Uses of alternative data sources for public health statistics and health policymaking: challenges and opportunities. *Joint Statistical Meetings*. Vancouver, BC.

[63] National Institutes of Health. (n.d.). NIH strategic plan for data science. Retrieved on August 29, 2018, from https://datascience.nih.gov/sites/default/files/NIH_Strategic_Plan_for_Data_Science_Final_508.pdf).

[64] National Academies of Sciences, Engineering, and Medicine. (2017). *Innovations in federal statistics: Combining data sources while protecting privacy and federal statistics, multiple data sources, and privacy protection: Next steps.* Washington, DC: The National Academies Press. doi: 10.17226/24652.

[65] Rothwell, C. (2018, August). NCHS director addresses challenges for federal statistical agencies. *AMSTATNEWS*. Retrieved on August 29, 2018, from http://magazine.amstat.org/blog/2018/08/01/nchsdirector/.

[66] Dept. of Homeland Security, Privacy Policy Guidance Memorandum 2008-01, "The Fair Information Practice Principles: Framework for Privacy Policy at the Department of Homeland Security." Memo from Hugo Teufel III, Chief Privacy Officer, Dec. 28, 2008, Retrieved on Jan. 22, 2019 from https://www.dhs.gov/xlibrary/assets/privacy/privacy_policyguide_2008-01.pdf.

[67] PII is defined by OMB as "information which can be used to distinguish or trace an individual's identity, such as their name, social security number, biometric records, etc. alone, or when combined with other personal or identifying information which is linked or linkable to a specific individual, such as date and place of birth, mother's maiden name, etc." OMB Memorandum No. 07-16. *Safeguarding against and responding to the breach of personally identifiable information*. May 22, 2007.

[68] 5 U.S.C. § 552a(e)(1).