DOCUMENTATION FOR THE GENERIC CLEARANCE FOR THE COLLECTION OF QUALITATIVE RESEARCH & ASSESSMENT

TITLE OF INFORMATION COLLECTION:

Big Data or Not: Determining the Use of Big Data

[X] INTERVIEWS
[ ] SMALL DISCUSSION GROUPS
[ ] FOCUS GROUPS
[ ] QUESTIONNAIRES
[ ] OTHER (EXPLAIN:)

DESCRIPTION OF THIS SPECIFIC COLLECTION

1. **Intended purpose**

The U.S. Department of Health and Human Services Office of the Assistant Secretary for Planning and Evaluation (HHS/ASPE) proposes to conduct interviews with subject matter experts in big data (BD) for the potential use of data to answer questions related to HHS programs. The subject matter experts will be from academic, government, and industry organizations.

2. **Need for the collection**

Frameworks for analyzing big data (BD) are still being developed and the overall scientific validity of such new approaches and data sources are not well understood. At present, information is scattered and anecdotal. There is a need for HHS to better understand the data and technologies that underlie BD and determine whether and how they might be incorporated to improve evidence-based decision-making by the Department. The information obtained by this project will help address these questions.

For the purpose of this project we will use the current definition of the BD found in the NIST Big Data Working Group draft (see https://bigdatawg.nist.gov/_uploadfiles/M0613_v1_3911475184.docx).
Big Data refers to the inability of traditional data architectures to efficiently handle new types of data that is generated by modern information systems. Examples of this data include personalized activity data captured through motion sensors, and syndromic surveillance data that might be captured through analyzing web queries and monitoring over-the-counter (OTC) medication purchases. Characteristics of Big Data that force new architectures are as follows:

Volume (i.e., the size of the dataset);
Variety (i.e., data from multiple repositories, domains, or types);
Velocity (i.e., rate of flow); and
Variability (i.e., the change in velocity or structure).

The above definition serves to differentiate BD technologies from many traditional data resources found at HHS. For example, HHS survey data does not meet this definition. Surveys may seem large, but they are generally not large by current BD standards, which refer to datasets that are vastly more complex. Once collected and interpreted, even a data set as large as the Decennial Census of Population does not change in content or format (i.e. it lacks velocity, variety, and variability). It is fixed and static until the next Census ten years later.

In contrast the literature databases maintained by the National Library of Medicine are perhaps HHS's best examples of BD as these data resources are continuously updated (velocity) with over 1 million publications per year[1] (volume) that represent new experimental findings across a multitude of fields that publish on different schedules (variability) from a large number of sources (variety).

Another BD example is the Medicare Claims Database. This is an administrative database used to carry out our insurance obligations. It is estimated to grow at between 1.5 and 2 billion claim records per year (volume),[2] updated continuously each day (velocity), with changes at random intervals with thousands of different possible diagnosis and procedure codes (variability), from thousands of different providers and specialties (variety).

This project is a first step in understanding the utility of BD sources maintained by outside entities, including commercial sources. Different BD sources can provide novel information for evidence development and may have advantages over data generated by HHS due to the speed of data collection, the potential for geographic granularity, and relative lack of measurement error. For example the Centers for Disease Control and Prevention's National Syndromic Surveillance Program integrates electronic health information from emergency departments, urgent care, ambulatory care, inpatient care, pharmacy data, and lab data, with standardized analytic tools to support detection of and rapid response to hazardous events and disease outbreaks.[3]

However, BD sources must be evaluated carefully to assure the data are of adequate quality. The strengths and weaknesses of different sources and kinds of data should inform how they are applied. The speed of data collection for some BD sources makes the data promising for real-time surveillance and monitoring and for generating hypotheses to be investigated using higher-quality data sources. For example, the FDA's Sentinel Initiative combines electronic health records, insurance claims data, and registries for adverse event monitoring to ensure safety of drugs and other regulated medical products.[4] A distributed data infrastructure allows for rapid analysis across the database of more than 193 million patients.[5] The use of the Common Data Model helps ensure standardization and maintain data quality. Methods are refined to get more precise estimates when an issue is detected, a process called "signal refinement." However, the same BD source may be less well-suited to other applications. For example, if the data is not representative of vulnerable populations, decisions based on that data may have unintentional discriminatory impacts. Health service researchers throughout HHS should understand when and

[1] *See* National Library of Medicine, "Yearly Citation Totals from 2017 MEDLINE/PubMed Baseline: 26,759,399 Citations Found," *available at* https://www.nlm.nih.gov/bsd/licensee/2017_stats/2017_Totals.html visited (Feb. 26, 2018).

[2] Office of Enterprise Data and Analytics, Centers for Medicare and Medicaid Services (personal communication, Feb 26, 2018).

[3] https://www.cdc.gov/nssp/index.html

[4] Robb, M.A., Racoosin, J.A., Sherman, R.E., Gross, T.P., Ball, R., Reichman, M.E., Midthun, K., and Woodcock, J. (2012). The U.S. Food and Drug Administration's Sentinel Initiative: expanding the horizons of medical product safety. *Pharmacoepidemiology and Drug Safety, 21*(S1), 9-11.

[5] Popovic, J.R. (2017). Distributed data networks: a blueprint for big data sharing and healthcare analytics. *Annals of the New York Academy of Sciences, 1387*(1), 105-111.

how to best employ BD analytic tools so that they may use the most appropriate data sources and analytic tools available to them in testing their hypotheses.

The information collection proposed in this project will provide an overview of the current state of the art of BD.

3. **Planned use of the data**

The information collected from this study will be used for identifying opportunities, challenges, and key issues in using BD at HHS.

4. **Date(s) and location(s)**

Data collection will begin in March 2018 and continue through May 2018.

5. **Collection procedures**

The project will consist of 21 to 25 semi-structured interviews with BD subject matter experts (SMEs), either over the phone or in-person, for approximately 60 minutes each. The contractor will contact those selected to schedule their interviews by first emailing them an invitation, followed by a second email to answer any questions they may have and to determine their willingness to be interviewed. After SMEs have agreed to participate in the study, then the interviews will be conducted, recording respondents' answers in writing and audio recording the sessions. Audio recordings of the interviews will be reviewed to update the written information so that we have a complete record of feedback from each interview.

6. **Number of collections (e.g., focus groups, surveys, sessions)**

There will be a maximum of 25 subject matter experts interviewed.

7. Description of respondents/participants

Interviewees will be individuals with extensive knowledge of big data used in their field of work.

8. Description of how results will be used

The results will be used to understand the technical, workflow, and business issues to consider in the use of BD at HHS.

9. Description of how results will or will not be disseminated and why or why not

Results from this project will be summarized in a final report by the contractor. The report will be posted on ASPE's website. Respondents will receive a summary of the interview for their review.

AMOUNT OF ANY PROPOSED STIPEND OR INCENTIVE

There is no stipend or incentive for participation.

BURDEN HOUR COMPUTATION *(Number of responses (X) estimated response or participation time in minutes (/60) = annual burden hours):*

Type of Respondent	No. of Respondents	No. of Responses per Respondent	Average Burden per Response (in hours)	Total Burden Hours	Hourly Wage Rate	Total Respondent Costs
Big Data Subject Matter Expert	25	1	1	25	$67.91	$1,697.75
TOTALS	25	1	1	25		$1,697.75

BURDEN COST COMPUTATION

Staff	Average Hours per Collection	Average Hourly Rate	Average Cost
Medical Officer GS 15	37.5	$77.58	$2909.25
Presidential Management Fellow GS 11	37.5	$31.87	$1195.13
Estimated Total Cost of Information Collection	$ 4104.38		

OTHER SUPPORTING INFORMATION

REQUESTED APPROVAL DATE: March, 2018

NAME OF CONTACT PERSON: James Sorace, Medical Officer, HHS/ASPE

TELEPHONE NUMBER: 202-205-8678

DEPARTMENT/OFFICE/BUREAU: U.S. Department of Health and Human Services (HHS), Assistant Secretary of Planning and Evaluation (ASPE)