Big Data or Not: Determining the Use of Big Data

ASPE Generic Information Collection Request
OMB No. 0990-0421

Supporting Statement – Section A

Submitted: March 21, 2018

Contracting Officer Representative
James Sorace MD MS
Medical Officer
U.S. Department of Health and Human Services
Office of the Assistant Secretary for Planning and Evaluation
200 Independence Avenue SW, Washington DC 20201
202-205-8678
James.sorace@hhs.gov

Section A – Justification

1. Circumstances Making the Collection of Information Necessary

The Office of The Assistant Secretary for Planning and Evaluation (ASPE) at the Department of Health and Human Services (HHS) has contracted NORC at the University of Chicago to conduct the project “Big Data or Not: Determining the Use of Big Data”. This project aims to create an information resource to facilitate HHS decision-making regarding the use of big data (BD) to address questions of interest in this field. Comparative methodological frameworks for analyzing big data 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, anecdotal, and not well-documented.

2. Purpose and Use of the Information Collection

There is a need for HHS to better understand the data and technologies that underlie BD and determine if and how they should be incorporated to improve evidence-based decision-making by the department. The information obtained by this project will help address this question.

For the purpose of this project we will use the current definition of the BD found in the National Institute of Standards and Technology (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 the new datasets. 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

[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).

intervals with thousands of different possible diagnosis and procedure codes (variability), from thousands of different providers and specialties (variety).

HHS is now in an environment in which it must evaluate the utility of BD sources maintained by outside entities, including commercial sources, to meet many aspects of its mission. Examples of these BD sources 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 medication purchases. 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. 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.[3] A distributed data infrastructure allows for rapid analysis across the database of more than 193 million patients.[4] 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 are not representative of vulnerable populations, decisions based on that data may have unintentional discriminatory impacts. It is critical that health service researchers throughout HHS understand when and 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 that will help guide further HHS research efforts in the field.

3. Use of Improved Information Technology and Burden Reduction

Information will be collected via telephone interviews. We will use audio recorders and computers to take notes and manually extract information and themes across interviews.

4. Efforts to Identify Duplication and Use of Similar Information

To our knowledge, there is no similar effort to collect this information using a similar methodology and directed at the use of big data specifically for purposes of HHS programs.

5. Impact on Small Businesses or Other Small Entities

None

[3] 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.

[4] 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.

6. Consequences of Collecting the Information Less Frequently

This request is for a one time data collection.

7. Special Circumstances Relating to the Guidelines of 5 CFR 1320.5

There are no special circumstances with this information collection package. This request fully complies with regulation 5 CFR 1320.5 and will be voluntary.

8. Comments in Response to the Federal Register Notice and Efforts to Consult Outside the Agency

This information collection is being conducted using the Generic Information Collection mechanism through ASPE-OMB No. 0990-0421.

9. Explanation of Any Payment or Gift to Respondents

No incentives will be provided in this study.

10. Assurance of Confidentiality Provided to Respondents

We are not asking any personally identifiable information of respondents, rather only about their experience in their professional capacity. We are asking them to provide information about their experiences and use of big data.

11. Justification for Sensitive Questions

We will not be asking any sensitive questions.

12. Estimates of Annualized Burden Hours and Costs

The SME interviews will be approximately 60 minutes to complete.

Table A-12: Estimated Annualized Burden Hours and Costs to Respondents

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	1.0	25	$67.91	$1,697.75
TOTALS	25	1	1.0	25		$1,697.75

13. Estimates of Other Total Annual Cost Burden to Respondents or Record Keepers

There will be no direct costs to the respondents other than their time to participate in the data collection.

14. Annualized Cost to the Government

Table A-14: Estimated Annualized Cost to the Federal Government

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

15. Explanation for Program Changes or Adjustments

This is a new data collection.

16. Plans for Tabulation and Publication and Project Time Schedule

The information shared by the subject matter experts will be collected via typed notes and audio recordings. After every three to four interviews are completed, the consultants will review and analyze the respondents' answers to the interview questions and also any questions they may have raised. Given the small number of interviews, manual coding and analysis may be more efficient than a qualitative data analysis software package. This process will be followed until all of the interviews are completed, and then NORC will begin its final analysis to the interview data.

Timeline:

Completion Date	**Major Tasks/Milestones**
November 2017	Plan recruitment of SMEs Develop interview protocol and data collection rubrics Develop SME interview list
March 2018	Finalize interview protocol and data collection rubrics Finalize SME interview list Submit request for OMB approval
May 2018 – July 2018	Conduct SME interviews Finalize interview notes Summarize interview themes
August 2018 – September 2018	Produce draft report Revise and finalize report

17. Reason(s) Display of OMB Expiration Date is Inappropriate

We are not requesting an exemption.

18. Exceptions to Certification for Paperwork Reduction Act Submissions

There are no exceptions for the certification. These activities comply with the requirements in 5 CFR 1320.9.

LIST OF ATTACHMENTS – Section A

A. Subject Matter Expert interview guide