NCVHS
National Committee on Vital and Health Statistics

March 20, 2014

The Honorable Kathleen Sebelius
Secretary
Department of Health and Human Services
200 Independence Avenue, S.W.
Washington, D.C. 20201

Re: Steps to Improve the Usability, Use and Usefulness of Selected Online HHS Data Resources

Dear Madam Secretary,

The National Committee on Vital and Health Statistics (NCVHS) serves the Department of Health and Human Service (HHS) as the statutory public advisory body on health data, statistics, and national health information policy. In 2009 President Obama's Open Government Initiative was launched with the goal of opening government data resources to the public in a format that is both secure and useable. In 2013 President Obama also signed an Executive Order requiring that going forward, government data disseminated to the public be open and machine readable, wherever possible and legally permissible. In addition, the Affordable Care Act authorized HHS to release new data resources that advance transparency in the health insurance and health care provider markets in significant ways. With this “data liberation” emphasis the expectation was that opportunities would emerge whereby users of HHS data could get the information they need-- doctors deliver better care, employers promote health and wellness, and local policymakers execute better-informed decisions. HHS has been a leader in this effort, with data sets available through *healthdata.gov* recently surpassing 1000 entries. HHS resources continue to grow and are found in various websites including Healthdata.gov and the Health Indicators Warehouse, among other locations. HHS also releases data and makes it available through a variety of other means.

In 2012, the Work Group on HHS Data Access and Use was created within NCVHS at the request of the Chief Technology Officer of HHS. The charge of the Work Group is to advise HHS with ideas and recommendations to promote and expand access to HHS online data and to explore innovative uses and applications of the HHS data to improve health and health care. The Work Group is comprised of members of NCVHS with expertise in standards, privacy, population health and quality. In addition, the Work Group has consultant members who are nationally recognized for their expertise in the development of innovative data applications for health and health care and in population health and health information technology. The consultant Work Group members bring a particularly valuable perspective as they are active users of government data.

During the last 18 months and in response to the Charge, the Work Group has reviewed available online HHS data, focusing primarily on Healthdata.gov, the Health Indicators Warehouse, and the Health System Measurement Project and evaluated the usability, use and usefulness of the on line data. Observations and recommendations have been developed through discussions within the Work Group, as well as through discussions with HHS data staff, public and private data aggregation organizations, and individuals from local communities who are seeking data to support local health initiatives. Final input has come from the full NCVHS. The observations and recommendations contained herein address three areas of focus:

1. Usability of the data: "User friendliness" of the platform
2. Use of the data: Data documentation and metadata
3. Usefulness of the data: Timeliness of the data

All of the ideas and recommendations outlined below are offered for HHS consideration as resources permit, to the extent that they are practicable and feasible within HHS programs and priorities, and as future upgrades and improvements are planned in the context of the data life cycle.

1. Usability of the data: "User friendliness" of the platform

OBSERVATION

The healthdata.gov platform has continued to grow, evolve and improve since its inception in 2012. However, users of the data have expressed challenges in navigating through the data sets. As stated in the HHS Open Government Plan, the goal of data liberation is to make data available for use by providers, community leaders, researchers, population health professionals and policy makers. These users may have very different skills and very different questions, and it is difficult to meet all needs and desires through a single website. Successful websites directed at sales often anticipate the needs of the customer and help guide them to not only the initial area of interest, but also to related areas that may be of potential interest. The success of this strategy may be reflected in the Alexa website rankings below.

Site	View frequency rank	Bounce rate	Pages visited	Time on site (min)
Google	1	17%	20	18
Amazon	5	28%	11	9
Healthdata.gov	161,924	49%	2.9	<2
Health Indicators Warehouse	3,344,689	NA	1.3	NA

Industry giants such as Google and Amazon have continuously improved their customer focus such that: 1. The content is of interest to the visitors and leads to many visits and revisits to their sites. 2. Once at the site, the visitor often finds what he/she is looking for, as evidenced by a low bounce rate. 3. Engagement is suggested by the number of pages viewed and the average time spent at each visit. While a consumer retail website is clearly not strictly comparable to a public data website, the numbers suggest that there may be learning opportunities for the federal websites. These two private companies' websites were not as intuitive at the start, but they have evolved over time and set a high bar of performance with the resultant high level of expectation for consumers when interacting on the web.

RECOMMENDATION

HHS open data websites could benefit from consideration of the following practices and functionalities in planning future upgrades and improvements to its open data websites;

- Borrow applicable best practices for the data platform from industry leaders (e.g., Amazon) in presenting web-based content.
- Create a feature which allows for user feedback about the use of the data set selected.
- Enhance applicability of the content highlighted by giving the user the option to identify his/her role: e.g., provider, employer, consumer, policy maker, etc. and direct the user to role-relevant resources.
- Provide users with a drill down menu on each choice that enumerates the key elements of a data set, e.g., format, inclusion of key designated elements, whether program is interactive, access to adjunctive data, interoperability, meta- data.
- Additional features that HHS could consider include:
    - A feature that provides users with "*users who used this, also looked at …*"

    - This might include other non-HHS data sets that include federal data that contains social and health-related information. This might also include a link to pre-built tools that are data/sector agnostic, such as Community Commons, Google Public Data Explorer or Tableau Public, ArcGIS Explorer that have broad reach and afford the opportunity to bring in data from other sources, e.g. other federal agencies.
  - Hyperlinks to sites where use of the data is demonstrated.
  - Customer ratings of the data set.
  - Functionality to visualize data in several formats (maps, graphs, tables.) Especially at the small community level, the effectiveness of the Open Government Initiative hinges not on people just being able to interpret the data, but on their being able to turn around and convince community leaders and stakeholders of the data's significance.

2. <u>Use of the Data: Data Documentation and Metadata</u>

OBSERVATION

Developers and users consistently expressed the need for better data documentation, including plain language descriptions of the data sets, and data definitions. There is also a need for uniform capture of data tags or metadata. *A metadata record is a file of different types of information, usually presented as an XML document, which captures the basic characteristics of a data or information resource. It represents the who, what, when, where, why and how of the resource" (Federal Geographic Data Council).* Data sets often do not provide enough information about the data (structural and descriptive metadata) for potential users to understand and use the data appropriately. Going forward, the Executive Order of May 9, 2013 entitled, "Making Open and Machine Readable the New Default for Government Information" establishes a framework to help institutionalize the principles of effective information management at each stage of the information life cycle. It requires agencies to collect or create information in a way that supports downstream information processing and dissemination activities, including machine readable and open formats, data standards and common core and extensible metadata for all new information creation and collection efforts.

However, the existing data sets, created before this mandate, present challenges as key metadata element and data definitions are not uniformly available. An illustrative example of current challenges is Medicare's "Provider of Services File – Other," which includes hundreds of variables about more than 142,000 U.S. health care providers. Fields names are not always intuitive. Descriptions of key variables are not consistently provided or do not contain sufficient information. Descriptions may provide how data are collected, but not why they were collected. Provider types are not listed; the key variables are not generally or specifically described. Data documentation is not easy to navigate and is often provided on several different webpages and in a few separate files.

RECOMMENDATION

HHS could facilitate the use of the online HHS data by providing more information about data sets in a clear format that is easily understood by a range of audiences.

- While there is an embedded tension between timely release of data and completeness of data description, data publishers should provide a data item definition in a machine-readable format with easy to use and comprehend variable names and plain language descriptions, to complement sector-specific terms, spelling out any acronyms and abbreviations.
- Data publishers should apply the elements of the Common Core Metadata Schema as defined by the Office of Management and Budget to existing data sets where practicable. [1]
- Data publishers should present this information as if the audience were not familiar with the data set, the data system, or the data collection in order to achieve usefulness for a wide range of audiences.
- Data that are not in a machine readable format (e.g. PDFs) should also have data descriptions.
- Data elements should have a taxonomy, entity relationship diagram, and relationships of attributes identified.
- Each topic should have a descriptor that indicates the general topic. In addition to facilitating locating relevant data, it would also assist in capturing the volume of data sets related to each topic, potentially pointing to underrepresented areas.

3. Usefulness of the Data: Timely Availability of the Data

OBSERVATION

Currently, on healthdata.gov, links to over 1000 HHS "data sets" are provided and available for public use. The currency and timeliness of the various data sets varies from current year to several years old. Stakeholders often need more timely data to support rapid cycle innovation and decision making. The timeliness of the publication of the data is a function of the life cycle of data planning, data collection, data analysis and dissemination, and not necessarily related to the criticality of the health information contained in

[1] "Official Supplemental Guidance on the Implementation of M-13-13 'Open Data Policy – Managing Information as an Asset'. Office of Management and Budget. Accessed March 3, 2014 http://project-open-data.github.io/schema/)].

the data set. Extended lag time between data collection and data release can diminish the value of the data and the opportunity for action. For example, mortality data may not be time-sensitive for known chronic conditions, but may be urgently needed for surveillance to understand an emerging new fatal illness. Vital records data currently compiled from states may lag for several years. For example, six out of seven mortality indicators are from 2010.[2] A related metadata issue is that in some data sets timeframes are complicated by comingling of the year of the content with the year of the collection and/or the point at which the data were refreshed.

HHS agencies have taken a number of steps to improve timeliness and shorten the data collection–data release schedule, including early releases that have reduced the lag to the current year or even six months and also through technological enhancements to speed up the data collection process. There are several examples of how data availability has been accelerated and these approaches may serve as templates for other HHS data sets. The National Center for Health Statistics (NCHS), working closely with the vital statistics jurisdictions and the National Association for Public Health Statistics and Information Systems (NAPHSIS), has developed three approaches to shorten lag time between reporting vital events to the states and final public release from NCHS: 1) expanding electronic data collection for birth and death records; 2) incentivizing jurisdictions using financial rewards to promote more timely reporting; and 3) expediting release of vital statistics.

Another approach has been to provide data that are nearly complete to use for modeling rather than waiting for 100 percent completion. Medicare claims data that HHS has made available for the Pioneer Accountable Care Organizations, have proven to be extremely useful for applying predictive analytics to better coordinate and care for target patient populations. The tradeoff has been that initial data contain only 33 percent of claims, but even that small sample has been sufficiently specific to address population health opportunities. Each month the data are updated and become more complete over time. When possible, and depending on the use and type of the data, this approach could be extended to more HHS or federal data sets to address the issue of timeliness. Of course, safeguards must be in place as it is imperative that acceleration of the timeliness does not have a negative impact on the representativeness or quality of released data.

RECOMMENDATION

[2] Centers for Disease Control and Prevention. "Sortable Stats Data Sources." Accessed February 18, 2014 http://wwwn.cdc.gov/sortablestats/Report_Docs/PDFDocs/Sortable_Stats_Data_Sources.pdf

The overarching recommendation is that HHS should continue to identify ways to accelerate the release of selected data sets whose relevance is time-sensitive. The Working Group recommends that HHS consider the following approaches to identify where earlier release is valuable and ways in which to expedite release:

- In collaboration with key stakeholders identify high-demand, high-value data sets whose timely release is critical to improving health and health care.
- Encourage data custodians to review the data life cycle of selected high value data sets to see if they can improve timeliness of segments of the life cycle.
- Encourage and support the use of technology in the data collection and processing stages in order to decrease processing delays.
- If appropriate, release early incomplete partial data sets marked as "provisional," and provide guidance on how data users should assess and interpret provisional data. Criteria should be developed for which data sets might be released as provisional.

The Committee is in the process of identifying a number of other areas of opportunity that will be the focus of future letters. As always, NCVHS stands ready to assist the Department in implementing these recommendations and further exploring key concepts and ideas.

Sincerely,
/s/

Larry A. Green, M.D. Chairperson,
National Committee on Vital and Health Statistics

Cc: HHS Data Council Co-Chairs