Good afternoon. My name is Robin Pendley. I am joined here by my co-presenter Kate Brett. Today, we will be talking about the open-data health initiative of the National Center for Health Statistics/Centers for Disease Control and Prevention called the Health Indicators Warehouse (HIW).
The United States Department of Health and Human Services (HHS) has historically held a vast amount of data that could potentially improve population health. The HIW was formed as a data hub for the HHS Community Health Data Initiative and remains a flagship initiative in the HHS Open Government Plan.

Our (HIW) objectives include: 1) to enhance the understanding of population health and health care system performance in communities and 2) to spark and facilitate action to improve performance and value.

Our initial approach involved: 1) catalyzing the evolving network of community health data suppliers, application developers, and end users; 2) recruiting early adopters across an array of potential uses; and 3) increasing awareness of community performance to facilitate and inform action.

HIW is maintained by the NCHS, one of 13 separate statistical agencies within the US federal government and the principle health statistics agency. HIW is a collaboration of many agencies and offices within HHS, who provide data, support, and funding.
Some of the key features and functions of the HIW include a structured database of high quality data and metadata. Methodology and data source details are provided for each indicator and links are provided for original data sources as well as evidence-based interventions (where possible). Geographical break-outs when available are provided. Data can be presented in tabular, chart, or map form.

HIW is meant to be an accessible interface for rapidly locating indicators, reviewing data, and downloading content in csv format. A key component of the HIW is the Application Programming Interface (API) web services that make the entire database open and machine readable.

The API access is designed to present information to systems with disparate architectures and technologies. We provide RESTful services and a Try It-REST Tester that includes a URI builder.
Currently there are over 1200 indicators in the HIW. These come from roughly 180 data sources including: federal and state governments, trade associations, and non-governmental organizations. Despite the large number of data sources and indicators there are only 20 different suppliers of the indicators. Data suppliers directly provide data to HIW staff such as NCHS. The data source is primary source of the data such as National Health and Nutrition Examination Survey.

The largest number of indicators come from two governmental initiatives. As of April 2015, Healthy People 2020, a national initiative to improve the health of Americans, provides 960 indicators, while the Centers for Medicare and Medicaid Services provides 166 unique indicators. These two indicator sources provide 92% of the indicators in the HIW, but less than 3% of the data, because there are generally not many estimates associated with each of these indicators. Some of the indicators only have a total and two age break-outs while available by several geographic levels, while others have only national estimates which may or may not be divided by demographics.

Tabulated public use indicators are obtained from many other direct sources, such as the Census Bureau. Roughly 50 indicators are obtained in this way.

Finally, some data is obtained directly from the source and are then released either uniquely to this project or are tabulated separately from their other dissemination methods. Generally, these are vital statistics indicators. A large proportion of the data
in the HIW is associated with this mechanism, due to the fact that we tabulate these indicators for differing numbers of grouped years, and for many different demographic categories.
This is a screenshot of an HIW data table. As you can see, we take our data as it is given to us and not all possible breakouts are available for all years. The uniform method of display is showing the demographic break-outs as rows, and years as columns. We present estimates and confidence intervals for each breakout. In the underlying data, but not shown in the interface, are estimates of the standard error for each estimate as well as the raw numerator and denominator.
For each indicator, we provide a substantial array of metadata information detailing its source and creation, and information to describe how it might be useful. This slide lists most of the important fields we include where appropriate and meaningful. Most indicators use most of the fields, but generally not all of them. We attempt to provide enough information for users to understand the data.
I am now going to talk about the issues we have encountered in starting up this dissemination tool and some of the solutions we have found. I will begin a series of challenges having to do with the data itself.

**Timeliness**

Because we get so many of our indicators from outside sources, we must constantly review websites to know when new data is released. The process for uploading data into the system is time consuming and so our current funding limits the number of annual data loads into the system. Finally, the two suppliers of a large number of indicators provide the metadata in a special format. Thus, two data loads each year are set aside to load them separately from other data.

To deal with these issues, we have worked closely with our contractor to make each data load more efficient and automated. One of the first process changes we made was to format all data records into a very specific XML format. This allowed the processes to run more smoothly. Second we group data from multiple indicators and data suppliers together wherever possible to use the limited number of data loads. Finally we accept that the data is not always up to date. Looking ahead, we are working to obtain data through API feeds as those systems come online and obtain notifications of data updates through automated systems as available.
The format data are obtained in is a continuing challenge. Each of our data suppliers provides their data in different formats. Furthermore, we find that the formats change over time from individual suppliers, so while there is some decrease in the amount of programming effort required over time, it is not eliminated. The reason is that the HIW cannot require source data to be in a particular format; we are offering a service to the data user but in most cases the supplier has not requested the inclusion of their indicators in the HIW.

We are addressing these issues by carefully allocating the programming of incoming data. We plan on a longer period of time being necessary to program indicators for the first time. We expect that the time invested will reduce programming effort in following years, since the format of most incoming indicators stay relatively stable over time. When the format changes over time, it is generally because new fields are added or new indicators are included in a single file; that requires much less work to address than the initial programming. We also have put effort into creating our data files of record to be exactly the same. The format of these files then allows the use of basically the same program to output the data into XML files for uploading.
One of the biggest issues early on was the level of effort expended reviewing incoming files for errors. Most of the data were not created by us, and the two main data suppliers were sending us data either straight from their data processor or before internal cleaning. The HIW staff was finding errors, but because it was not our data, it was very difficult to know how to thoroughly identify all problems. In addition, we tried to use one standard set of geographical boundaries, even though at the sub-state county level, changes have happened over time and different data sources have incorporated these changes either at different times or even not at all.

To alleviate this problem, we have made it clear to all indicator suppliers that we load the data that is given to us, and it is up to them to make sure the data are correct. We validate estimates produced internally by our staff. Otherwise, other than checking for gross errors such as missing codes or percents greater than 100, we do not clean data. By accepting data that suppliers have confirmed meet their quality control requirements we are better able to use our limited staff resources.
Originally, part of the HIW mission was to create harmonized indicators, where each concept was measured using a single indicator. Over time, the difficulty of this task has become apparent. Some differences between indicators of the same concept were methodological. There are differences in the population sampled, in question wording, and even where within the questionnaire the concept was measured.

Another major source of discordance is in data processing. Many indicators are age-adjusted and how that process is carried out changes the resultant estimate. There are differences in decisions on how to deal with missing values, such as imputation versus dropping missings from the estimation process. Finally, there are numerous indicators that are collected by different agencies and each considers its own as the best. Given these difficulties, we have shifted our emphasis towards presenting multiple high-value indicators of the same concept, making sure that the labels and methodology notes are clear and complete. We encourage users to understand the differences and then choose which indicator to use.
Moving on to our metadata challenges, we have some issues which parallel our data issues. Timeliness is one such issue. With more than 1200 indicators, metadata is difficult to keep updated.

The fields that are most problematic are the references, the associated interventions, which were included as examples of accepted methods for creating change in the area of health being measured, and data source descriptions, including maintaining updated information on questionnaire wording as appropriate. The HIW staff are making progress on this issue but has come to accept the initial metadata input may remain and we work to refine and correct them as time allows.
The format we obtain the metadata in has become one of our major current issues. The HIW initially created metadata in simple spreadsheets, and one of the major data suppliers continues to use that format. We then created a relational data base in MS Access, which was adopted by the other major indicator supplier. The current database being used by the HIW is a SQL server database with a web application interface. We can edit and create indicator metadata through our interface but do not allow others directly into the system.

Loading the outside data is difficult. In the case of a spreadsheet load, we have experienced such issues as truncation by maximum characters and problems trying to load a flat file into a relational database. Loading from the outside relational database should be fairly straight-forward except that the data supplier continues to modify that database for their own use, and the modification have not been well documented.

One step we have found to be beneficial is to limit each data load to include only one indicator supplier that has their own metadata. The other method for addressing this challenge is spending a great deal of effort reviewing metadata once it is loaded into our system. We plan in the future to automate the metadata loads, which should further reduce both effort expended and errors introduced.
Consistency of the content in the metadata has been a concern of the HIW staff. There are multiple metadata suppliers, and within each supplier, multiple people creating and editing metadata. This raises issues on how to enforce rules on the content of each metadata field, such as how many and what keywords and topic areas should be used, what types of references should be incorporated, and how detailed the methodology notes should be.

The way we are moving forward is by taking the initial metadata template, creating rules for each of the fields, as best as we can describe them, and then internally spending time refining the metadata to make entries more consistent.
In this last slide, I bring up our most vexing problem, that of metadata look-up tables. In our metadata database, some of the fields are stored as look-up tables, to ease maintenance. This is helpful in maintaining fields that change fairly frequently and/or are used in numerous indicators.

Each supplier is maintaining their own metadata. The one main outside supplier using MS Access fortunately is in our same organizational unit and coordinates work with us. This is important because when outside metadata files are loaded, they overwrite whatever is currently in the main SQL database.

One solution to the problem is to provide separate ranges of number for each duplicate look-up table, so that edits to one will not overwrite data on the master. When there are important changes, we have coordinated efforts and made sure that both systems are edited simultaneously.

Finally, once we load metadata, we then push the look-up tables back to the outside groups for their use. In this way, while we potentially have multiple IDs with the same information, that issue is reduced as much as is practical while allowing both group absolute ownership of their own system.
In closing, the HIW was created as a resource for health and health-related data from across federal and non-federal data systems. We have created the system to allow the underlying data to be obtained through an API and in this manner serves the needs of the developers’ community. We also provide traditional data file downloads in CSV format for other users. Over time, we have come to realize that the maintenance of the system, both metadata and data, is challenging because of the fact that we obtain indicators from many different sources. The good news is that these issues are becoming more defined and known, so that there are fewer surprise problems to address.
Questions?

For Further Information:
HIW Website: http://www.healthindicators.gov/

Kate Brett
Captain, U.S. Public Health Service, National Center for Health Statistics
kbrett@cdc.gov

James Craver
Assistant Director, Office of Analysis and Epidemiology
Acting Branch Chief, Infant, Child, and Women’s Health Statistics Branch
National Center for Health Statistics
jcraver@cdc.gov

Robin Pendley
Senior Service Fellow, National Center for Health Statistics
rpendley@cdc.gov