The NCVHS Health Data Framework

Executive Summary

The benefits of the explosion and liberation of health related data can only be realized if the systems for making sense of data keep pace with their burgeoning volume and complexity. Otherwise we run the risk of being unable to efficiently analyze the data or to even compare them to one another because of their quantity and variety. The promise of data-driven progress toward the health system's triple aim will remain elusive unless we find ways to overcome this risk.

The National Committee on Vital and Health Statistics (NCVHS) Health Data Framework seeks to address this risk by facilitating dataset classification, use, and analysis. People in the health data world, as in others, sometimes mean quite different things by the same terms without being aware of these differences.[1] This has been called the Tower of Babel Problem. The work on the NCVHS Health Data Framework has revealed that perspectives, which may seem the same on the surface can turn out, on further examination, to be quite different. A major purpose of the evolving Framework is to create a "cross-walk" among vocabularies that will make it possible for everyone to understand one another when talking about and working with data.

NCVHS drafted two resources, a Data Structure and Methods Taxonomy to seed development of the Health Data Framework. These drafts offer a systematic approach to thinking, talking, and acting with respect to data. These resources also propose metadata to tag datasets to support re-use. The target audience for these resources includes statistical and analytic experts, researchers, data suppliers and intermediaries, and application or system developers.

Two complementary and linked resources together compose the Health Data Framework:

1. The **Data Structure**, a multi-dimensional structure for organizing data about populations at different levels or scales; and
2. The **Methods Taxonomy**, a taxonomy of dataset and secondary use characteristics, analytic and visualization techniques, stewardship principles and standards, to guide data use and re-use.

The Health Data Framework has three goals:

1. To help data experts support the health ecosystem to systematically use data from all relevant sources to solve problems;
2. To surface high-impact gaps in data sources and methods; and
3. To catalyze development of interactive tools to support optimal data use and learning.

[1] Petrie H, Do You See What I See? The Epistemology of Interdisciplinary Inquiry. Journal of Aesthetic Education, Vol. 10, No. 1 (Jan., 1976), pp. 29-43. University of Illinois Press. http://www.jstor.org/stable/3332007

Consider a futuristic scenario based on NCVHS's multi-year effort to understand how communities can become learning systems for health, and how their data use capacities can be enhanced.

> A community coalition has targeted childhood obesity as its top priority after talking with community members, and analyzing data on health disparities and assets. As they explore national survey data from their county, they link to a dataset aggregated by the State Health Department at the census tract level to identify a hotspot (high incidence and prevalence of childhood obesity) and the nearest coldspot (low incidence and prevalence). Then they link to a neighborhood dataset showing what community organizations these two areas have in common. Next they pull in a dataset of the school programs for the two areas. Their analytic workbench mediates authorization and access to the datasets, unpacks the data, and creates a display appropriate to the dataset and the question they are asking.[2] To be continued.

Although the Data Structure and Methods Taxonomy are described separately here, their chief value lies in the ways they interact and function together to guide dataset use to answer a specific community's health question and guide an intervention. For example, a dataset may be tagged with the metadata of its location within the Data Structure. The Methods Taxonomy provides additional metadata to tag datasets and methods, clarifying those which work together and where they apply in the Data Structure. Together, they organize information about what types of data are needed and available, and what methods for accessing, analyzing, linking, etc. are appropriate for each source.

If further developed, the Health Data Framework will provide a systematic way of determining how to collect and protect individual data under different circumstances, depending on the purpose. Anticipated benefits of the Framework are that it will:

- provide metadata to annotate datasets to clarify appropriate uses and identify limits to usability for a proposed secondary use
- make it possible to develop interactive tools to represent the view of the data supplier and put filters on the data that are appropriate to the purpose and adhere to stewardship principles
- compare techniques for re-purposing the data
- match semantic standards and versions used in the dataset
- disseminate stewardship principles.

It is anticipated that the Health Data Framework would serve as a filter to enable work at the appropriate population level, given the balance between the analysis required and the sensitivity and risk associated with using the data. Thus, it is envisioned the Framework would provide a way to control and structure the process.

1. Introduction

The benefits of the explosion and liberation of health related data can only be realized if the systems for making sense of data keep pace with their burgeoning volume and complexity.

[2]This scenario is based on the Use Case described in Appendix 2

Otherwise we run the risk of being unable to efficiently analyze the data or to even compare them to one another because of their quantity and variety. The promise of data-driven progress toward the health system's triple aim will remain elusive unless we find ways to overcome this risk.

The National Committee on Vital and Health Statistics (NCVHS) advises the Department of Health and Human Services on health data, statistics, privacy, and national information policy. The NCVHS Health Data Framework seeks to address this risk by facilitating dataset classification, use, and analysis. People in the health data world, as in others, sometimes mean quite different things by the same terms without being aware of these differences.[3] This has been called the Tower of Babel Problem. The work on the NCVHS Health Data Framework has revealed that perspectives, which may seem the same on the surface can turn out, on further examination, to be quite different. A major purpose of the evolving Framework is to create a "cross-walk" among vocabularies that will make it possible for everyone to understand one another when talking about and working with data**.**

NCVHS drafted two resources, a Data Structure and Methods Taxonomy, to seed development of the Health Data Framework. These drafts propose a systematic approach to thinking, talking, and acting with respect to data. These resources also propose metadata to tag datasets to support re-use. This is akin to NLM indexing of journals and tagging of articles so that they can readily and systematically be searched to inform clinical and public questions. It is key to being able to find a particular article or to structure a systematic review of the literature. In either case, the function of tagging studies and articles reaps many-fold the investments of NIH and federal agencies.

NCVHS is engaged in a multi-year effort to understand how communities can become learning systems for health, and how their data use capacities can be enhanced.[4] Examples from these roundtables guided initial development of the Health Data Framework.

The Health Data Framework Project has three goals:

1. To help data experts support the health ecosystem to systematically use data from all relevant sources to solve problems;
2. To surface high-impact gaps in data sources and methods; and
3. To catalyze development of interactive tools to support optimal data use and learning.

With this white paper, NCVHS seeks to inspire the Federal government and the data supplier ecosystem to elaborate the Health Data Framework, which is described below. The Committee

[3] Petrie H, Do You See What I See? The Epistemology of Interdisciplinary Inquiry. Journal of Aesthetic Education, Vol. 10, No. 1 (Jan., 1976), pp. 29-43. University of Illinois Press. http://www.jstor.org/stable/3332007

[4] NCVHS serves as the statutory (42U.S.C.242k[k]) public advisory body to the Secretary of Health and Human Services on health data and statistics. In that capacity, it provides advice and assistance to the Department and serves as a forum for interaction with interested private sector groups on key issues related to population health, standards, privacy and confidentiality, quality, and data access and use. Its 18 members have distinction in such fields as health statistics, electronic interchange of health care information, privacy and security of electronic information, population-based public health, purchasing or financing health care services, integrated computerized health information systems, health services research, consumer interests in health information, health data standards, epidemiology, and the provision of health services. http://ncvhs.hhs.gov/

hopes to stimulate an ongoing dialogue that further develops resources for communities and other data users.

The contents of the white paper are as follows:

- Section 1. Introduction
- Section 2. Overview of the Health Data Framework
- Section 3. Topics and Issues of Interest
- Section 4. Paths Forward and Vision of the End Game

Enabling Communities to Become Learning Systems for Health[5]

First, let us consider why the Health Data Framework is needed. America's communities face a growing set of pressures to use data effectively in their local health improvement efforts. Some communities are seeking to brand themselves by their healthy lifestyles. Many are tackling pressing community problems such as teen pregnancy and drug overdose. The vigorous Federal data liberation initiative is rapidly increasing data access. There are new forms of accountability for non-profit hospitals and public health departments, and incentives to share data for collective impact. A network of supportive organizations and websites offers a rich array of data and support. This confluence of forces gives communities ever-increasing prospects for leveraging data to better understand and improve community health.

Despite these influences, many communities lack the capacity to take advantage of the expanding resources. Most data users may be unaware of sources outside their own arena (health care, public health, education, the private sector, and so on); or they may be aware that other data exist but not know how to analyze data from multiple sources. Perhaps they work with data at a single level of aggregation (individual, healthcare catchment area, county population) and don't know how to move among several levels, or how to look at data on upstream determinants such as economic resources or the built environment in conjunction with data on health outcomes, or how to choose the best data for evaluating the impact of interventions. The realities of non-interoperable data, data gaps, lag times, and uneven data quality, plus the shortage of local analysts, can add challenges to these already complex tasks. When data are brought together across perspectives, levels, and sources, the complexities multiply. And all these challenges are compounded by the absence of a common language that would enable effective communication about health data and methods.

The optimal use of data for community health requires extensive skills including systematically locating relevant available data; interpreting standards and applying principles of data stewardship for using multiple types, levels, and sources of data; identifying data gaps and designing strategies for filling them; and understanding the appropriate uses and limitations of the data. To access such skills, communities need multi-dimensional partnerships and a supportive national infrastructure to turn to for support.

NCVHS believes that without appropriate systems and resources, even sophisticated communities could be overwhelmed by the pace and volume of data release and the complexities of using the data. The Health Data Framework will be as complex as the datasets

[5] NOTE: Here or elsewhere, reference *'The Community as a Learning System for Health'* and note that this project helps to fulfill several suggestions for federal action, notably facilitating "the development and adoption of a national common reference information model for public health…." (page 31).

and techniques it describes. Its target audience includes statistical and analytic experts, researchers, data suppliers and intermediaries, and application or system developers. These experts, and the systems they develop, will use its classification resources to help community stake holders have simpler and more meaningful conversations as they work with data. NCVHS is eager to work with colleagues in the fields of public and community health, health care, and informatics to develop a common vocabulary and related resources that can serve as part of a supportive national infrastructure.

2. Overview of the Health Data Framework

NCVHS drafted two complementary and linked resources that together compose the Health Data Framework:

1) The **Data Structure**, a multi-dimensional structure for organizing data about populations at different levels or scales; and

2) The **Methods Taxonomy**, a taxonomy of data set and secondary use characteristics, analytic and visualization techniques, stewardship principles and standards, to guide data use and re-use.

Although the Data Structure and Methods Taxonomy are described separately here, their chief value lies in the ways they will interact and function together to guide data use. For example, a data set may be tagged with the metadata of its location within the Data Structure. The Methods Taxonomy will provide additional metadata to tag datasets and methods, clarifying those which work together and where they apply in the Data Structure. Together, they organize information about what types of data are needed and available, and what methods for accessing, analyzing, linking, etc. are appropriate for each source.

Consider the following scenario[6]: A community coalition has targeted childhood obesity as its top priority after talking with community members, and analyzing data on health disparities and assets. An analysis by the State Department of Public Health for their county, aggregated at the census tract level, show marked disparities in the incidence and prevalence of obesity in different areas and populations. Coalition members discuss what data they can marshal to guide decisions about target populations, interventions, outcome measures, and so on. An urban planner mentions a geocoded dataset with locations of bike trails, walking and other recreation resources. A school board member describes a dataset for the catchment area of each school, noting their programs related to nutrition and exercise. A school nurse mentions a dataset their school keeps with height and weight of students receiving insulin while at school. The coalition explores how these datasets relate to one another geographically—both overlaps and gaps—by placing them within the Data Structure. The Methods Taxonomy in turn helps them understand the privacy related restrictions that apply to the height and weight data on individual students. To be continued.

[6] This scenario is based on the Use Case described in Appendix 2

The Health Data Framework will provide a systematic way of determining how to collect and protect individual data under different circumstances, depending on the purpose. If fully developed, it could be used to:

- provide metadata to annotate datasets to clarify appropriate uses and identify limits to usability for a proposed secondary use
- make it possible to develop interactive tools to represent the view of the data supplier and put filters on the data that are appropriate to the purpose and adhere to stewardship principles
- compare techniques for re-purposing the data
- identify semantic standard and version used in the dataset
- disseminate stewardship principles

It is anticipated that the Health Data Framework would serve as a filter to enable work at the appropriate population level, given the balance between the analysis required and the sensitivity and risk associated with using the data. Thus, it is envisioned the Framework would provide a way to structure and organize the process.

Data Structure

The Data Structure is a multidimensional picture of the data space about populations at different levels. Exhibit 1 shows three of many dimensions. The dimensions represent different social-structural-biological variables such as the geographic scopes of populations, population health measurements, and the determinants of health.

Upper levels of a dimension are more general than lower levels. In the case of the geographic dimension addresses aggregated into census tracts, neighborhoods, and well defined civil divisions. Upper levels are not mere roll-ups of lower levels. For example, neighborhoods may include parts of multiple census tracts and cross sub-state and state boundaries.

Along the population health measurement dimension, lower levels are more proximal to the individual (person or intervention) and upper are summative for the population. For example, determinants provide a way to describe or explain a specific condition, while distal outcomes such as health related quality of life represent the collective impact of lower levels. As depicted in Exhibit 1, determinants are also a dimension of the Data Structure, ranging from proximal individual factors through social connections and living conditions to distal factors such as policy that provide context shaping the individual. Neighborhood compositional (e.g. median household income) and contextual (e.g. open spaces) factors are on this dimension as determinants. Measures of these determinants may be collected or aggregated at various levels along the geographic dimension (e.g. census tract, neighborhood, sub-state civil division, etc.)

Each dimension is a continuum —— that is, the boundaries of the cells are not fixed. Subdivisions may be added to clarify distinctions, or removed if a division is misleading. This picture of the way data from various levels fit together helps to systematically identify gaps in data sources and point to methods and strategies for filling the gaps, while applying relevant standards and stewardship principles.

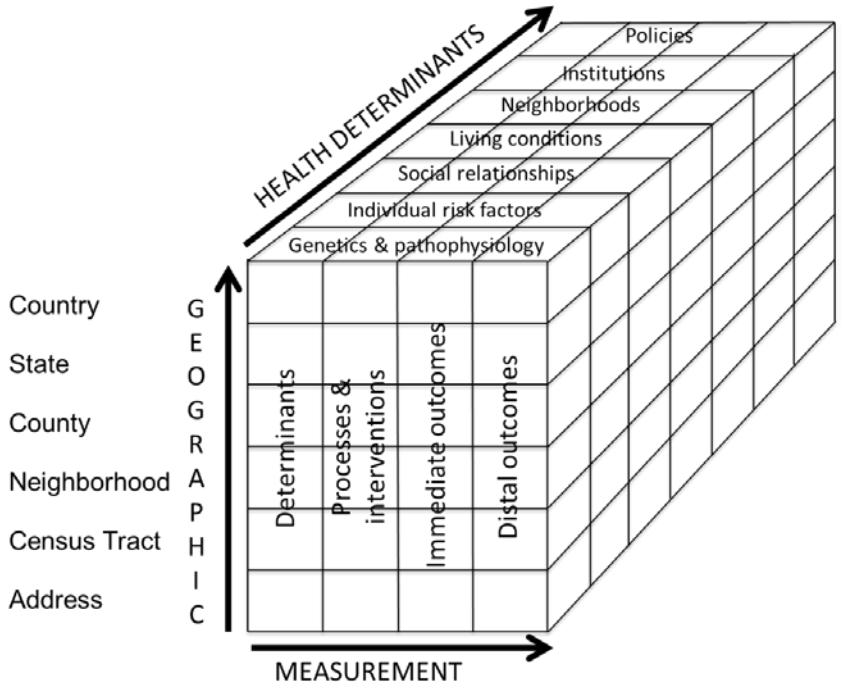


Exhibit 1.

Representation of the Data Structure

To reiterate, the Data Structure describes the hypothetical *data space,* not merely the data we have. This orderly depiction shows how particular data relate to other categories of data in the same dimension while retaining their distinctive characteristics and/or zones.

Examples of particular data:

- A state program of screening newborns for genetic abnormalities would be located in the State/Process & intervention/Genetics & pathophysiology cell.
- An intervention to reduce environmental asthma triggers in an apartment building would be located in the Address/Process & intervention/Living conditions cell.
- An individual's smoking status would be located in Address/Determinant/Individual risk factor cell; exposure to second hand smoke would be in the Address/Determinant/Living conditions cell.

Placing the data we have into the relevant data space(s) would generate a multi-dimensional "map" showing both the data we have and the data we don't (yet) have, thus providing a context in which to identify gaps[7]. In addition, the location of a dataset along applicable dimensions, or the coordinates that place it along multiple dimensions, can be used as metadata to tag that dataset to the Data Structure. As already noted, while the graphic in Exhibit 1 above illustrates the Data Structure using three dimensions, the Data Structure actually includes several other dimensions.

Exhibit 2 shows a broader range of dimensions of the Data Structure, plus the possibility of identifying new dimensions as the Framework is elaborated further.

[7] A gap does not need to be filled unless it is important to the analysis.

Exhibit 2. Data Structure

1. Geographic Dimension	a. Address
	b. Census tract
	c. Neighborhood[8]
	d. County
	e. State
	f. Country
2. Organization Dimension	a. Single
	b. Aggregate (roll-up)
	c. Virtual
3. Population Health Measurement Dimension[9]	a. Determinants
	b. Processes and interventions
	c. Intermediate outcomes
	d. Distal[10] outcomes: i. Disease specific scales ii. Health related quality of life iii. Summative (Health adjusted life years)
4. Determinants of Health Dimension	a. Genetic & constitutional pathways
	b. Pathophysiologic pathways
	c. Individual risk factors
	d. Social relationships
	e. Living conditions
	f. Neighborhood[11] compositional and or textual factors
	g. Institutions
	h. Social & economic policies
5. Pathophysiology Dimension	a. Risk factors for development of disease
	b. Asymptomatic primary pathophysiology
	c. Symptomatic primary pathophysiology
	d. Asymptomatic secondary pathophysiology
	e. Symptomatic secondary pathophysiology
6. Additional Dimensions	TBD
	TBD
	TBD

[8] On the geographic dimension, neighborhood is a level of data collection or aggregation.
[9] Institute of Medicine, *For the Public's Health* – measurement
[10] Proximal outcomes include process outcomes, e.g. (# of times a process is performed)/(# of opportunities to perform the process)
[11] On the determinants of health dimension, neighborhood compositional and contextual factors are determinants.

Methods Taxonomy

The Methods Taxonomy is a taxonomy of data set and secondary use characteristics, analytic and visualization techniques, stewardship principles, and standards to guide data use and reuse. Its categories can be used as metadata to tag datasets and methods, clarifying those that work together and where they apply in the Data Structure. This information can be used to document the biases of data and show how to use and repurpose the data.

Exhibit 3 presents a simplified version of the Methods Taxonomy, highlighting the first two of its many categories, and just the first of many levels of sub-categories.

Exhibit 3. Methods Taxonomy

1. Data Source Characteristics	a. Type of data source
	b. Original collector and aggregator
	c. Purpose of collection
	d. Method of collection
	e. Voice
	f. Granularity
	g. Primary users
	h. Primary uses
	i. Applicable regulations
	j. Identification status
	k. Consent provided at the time of data collection
	l. Applicable standards
	m. Demographic representation
	n. Vulnerable populations included
	o. Population health measures included
	p. Timing
	q. Accuracy
	r. Completeness
	s. Timeliness
	t. Limitations
	u. Biases
2. Secondary Data Use Characteristics	a. Users
	b. Uses
	c. Granularity
	d. Timing
	e. Timeliness
	f. (Etc.)

The data source characteristics in the Methods Taxonomy correspond to questions about a data set whose answers show how, or if, it can be reused. Similarly, the secondary use characteristics correspond to questions about a proposed secondary use whose answers show which data set characteristics are required and use limitations.

> **Example of Methods Taxonomy Use:** Under federal law and regulations, a health care provider may collect an individual's social and behavioral determinants, provided the data are to be used for a purpose related to the patient's health. In that case, identified data are *protected health information* (PHI). Federal law (HIPAA) allows the provider to disclose PHI to other health care providers and to a legally defined public health authority or for law enforcement, among other defined recipients. By tagging a data set with the category of original collector (health care provider), the purpose (individual's health), the identification status (identified), and also tagging the disclosure (secondary use) with the category of secondary user (public health authority) and use (public health), the combination of tags provides the metadata needed for future data users to systematically comply with the law and regulations.

The Methods Taxonomy is extensible —that is it takes future growth into consideration. A fuller version, with five categories and four levels of subcategories, is presented in Appendix 1. The subcategory levels can be elaborated. For example, for identification status, anonymized data can be decomposed into no linkage possible; re-linkable data; and linked with a protected key. Similarly, for consent provided at the time of data collection, consent by the individual can be decomposed into broad and unspecified; time limited consent; consented for partial, source specific use; and consent for the particular type of use.

The Data Structure and the Methods Taxonomy Work Together:

Although the Data Structure and Methods Taxonomy are described separately here, their chief value lies in the ways they will interact and function together to guide data use. Once it is developed, the Health Data Framework will be used to classify specific data sets to dimensions in the Data Structure and to subcategories in the Methods Taxonomy. In other words, the Framework will provide metadata to annotate data sets to clarify appropriate uses and identify limits to usability for a proposed secondary use; compare techniques for re-purposing the data; consider relevant stewardship principles; and unpack the data with the correct version of standards.

The Data Structure and Methods Taxonomy have many dimensions and subcategories to clarify specific relationships and differences. Only a subset will be applicable to a specific data set or analysis. When a category in the taxonomy is applicable, its subcategories provide standard terms for the classification.

Continuing the scenario of the community coalition targeting childhood obesity:[12] The coalition's initial review of availability of recreational resources shows that less than 1/3 of the obese students have such resources close to home. They notice others have access close to their school and others have access en route. They ask how much of their student population would be covered by a program that included these two types of access to recreational resources.

They browse the Data Structure and decide they are interested in data aggregated at the census tract level. They browse the Methods Taxonomy and note the elements that apply to their question. For example, under secondary use characteristics, they pick analysis of access to recreational resources for purpose, census tract for granularity of aggregation, and public health for use. Under analytic and visualization techniques, they drill down into modeling for type of analysis, and see a subcategory for techniques that handle multiple addresses (in this case the student's home, and their school). To be continued.

3. Topics and Issues of Interest

The Framework development process has already stimulated discussion of a number of topics and issues, some of which will need to be resolved in the future. Several are summarized below.

Outcome Data

Views differ about the best way to represent outcome data in the Data Structure. While the figure in Exhibit 1 shows intermediate and distal outcomes, some participants in the development process have argued that to be consistent with Donabedian's framework, outcomes should be represented in a single column.[13] Others have countered that intermediate outcomes warrant an independent column because that is the space in which government does much of its work. The latter group also points out communities need to look at intermediate outcomes to know if their interventions are having any effect on targeted aspects of community health. Informants agree that however they are sliced in the model, in reality outcomes exist on a continuum.

Organization and Geography

Perspectives vary on the question of how to represent organizations in the Data Structure — as a dimension on par with geography and population health measures, or as a sub-level within the geography dimension. Schools and health care organizations served as examples in this discussion. Those using a large integrated health plan as the paradigm favor embedding *geography* within *organization* because that type of organization has a geographic dimension. However, others argue for the importance of being able to independently vary organization and

[12] Scenario begins on page 5 and is based on the Use Case in Appendix 2

[13] According to Donabedian's model, information about quality of care can be drawn from three categories: "structure," "process," and "outcomes". Donabedian, A. (1988). "The quality of care: How can it be assessed?". *JAMA* **121** (11): 1145–1150.

geography because multiple competing large systems may serve the same geography. Participants have agreed that three levels of organization should be identified — single, aggregate, and virtual exhibit.

Levels of Data Collection and Aggregation

There is an important distinction between the level at which data are collected and the level at which they are aggregated and made available. Exhibit 4 depicts this distinction as a matrix, with the level of data collection on one axis and the level of aggregation on the other. These levels could be any of the levels along the geographic dimension in the data structure. The level of the community engaged in the analysis is the collection level — Sub-community are levels below that level and External are levels above. The collectors and aggregators could be private or public (governmental), and community groups or institutions might contribute their information to a virtual database made accessible to others.

Exhibit 4. Levels of Data Collection and Aggregation, with examples in cells

	COLLECTION LEVEL		
AGGREGATION LEVEL	**Sub-Community**	**Community**	**External**
Community	Data collected by individual schools & reported to the district	Police department's crime data[14]	National[15] survey
Sub-Community	A school's internal data	Police department's crime data reported by neighborhood	Health Department's report of sociodemographic characteristics by census tract or block group

Engaging the Community in Making Meaning from Data

Data collection and data analysis are linked by *purpose*, which determines the community's comfort level with various forms of analysis. It is also important to consider not only statistical significance but also meaningfulness. A community's decisions about data collection and analysis are filtered through the community's values, judgments and other priorities along with a sense of what people feel can be accomplished. Thus the design of data collection and analysis must be worked out not by analysts alone, but in substantive conversations with and among community members. They also have an important role in understanding what the outputs mean, and how to prioritize them.

[14] A town's police department, a county's sheriff department, and state police will all have data that may be relevant to an analysis.

[15] May include any survey above the level of the community on the geographic dimension, e.g. state, county, etc.

Timeliness

The issues of timeliness and granularity, particularly as they affect communities, surface frequently in NCVHS discussions. These are only two of several inter-related characteristics pertaining to data usefulness, along with accuracy, sensitivity and completeness. NCVHS addressed the issue of data timeliness in a March 2014 letter to the Secretary that presented observations and recommendations developed by its Working Group on HHS Data Access and Use.[16] Working Group members have suggested data sets be tagged with metadata describing timeliness, and also that data may be "fit to use" for some purposes before they are adequate for others. Following that logic, every data set could be tagged with metadata describing its timeliness, and uses could be tagged with metadata describing the timeliness required for each use. The tags on available data sets could be matched to the tags on the proposed use to determine when the data set was ready for the proposed use.

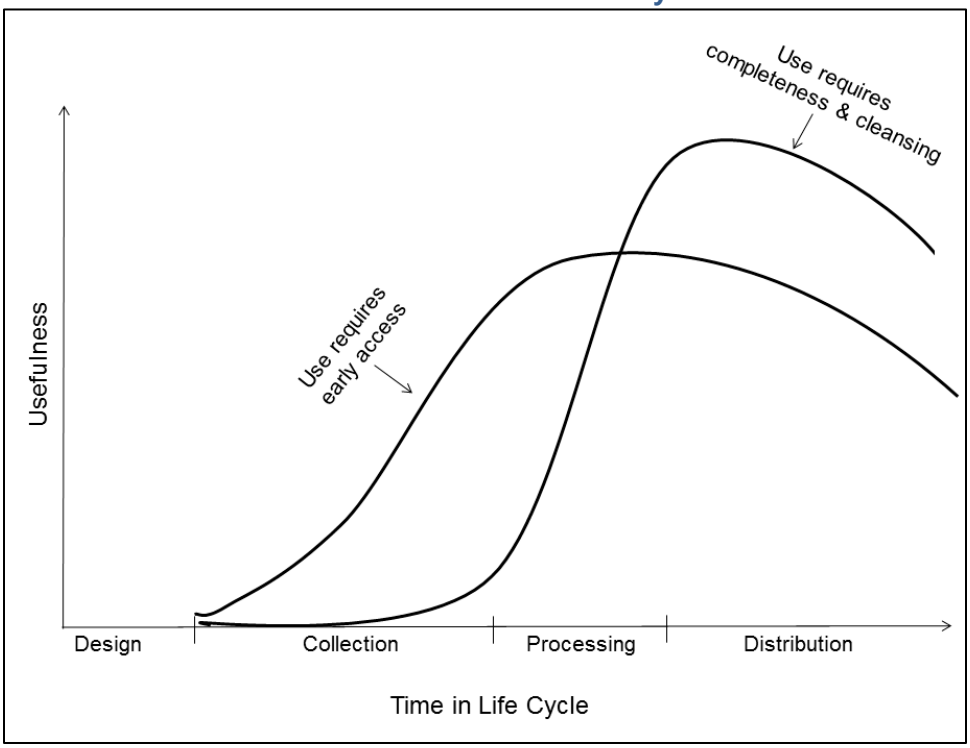
The attributes of data timeliness that are relevant in judging the fitness of data for specific uses include:

- Rate of change (how frequently to measure the subject of the data)
- Shelf-life of the data (how long the data are good for the intended purpose)
- Lag-time for validity (how long it takes for the data to become good)
- Acuity of need for the data (a major event, e.g. a change such as an increase in access to coverage with implementation of Medicare, starting a new cycle of data collection)
- Background rate of change (secular trends that contextualize the significance of the data)

Exhibit 5 shows a "timeliness lifecycle" that differentiates the concepts of shelf-life and lag-time, with corresponding increases and decreases in the usefulness of data. The notion of "fit for use" points to the diminishing value (for some cases or purposes) of waiting for data completion.

[16] http://www.ncvhs.hhs.gov/wp-content/uploads/2014/05/140320lt.pdf

Exhibit 5. Data Timeliness Lifecycle



Timeliness is related to other characteristics including timing (e.g., cross-sectional, longitudinal) and aggregator's judgment of fitness (e.g., provisional, open, closed).

Granularity

Community leaders raised the issue of data granularity at a 2011 NCVHS workshop on communities as learning systems for health.[17] One concern, for example, is that data aggregated and made available at the county level may hide important small area variation in social and health disparities. The centrality of neighborhood-level and small population-level information for meaningfully addressing community health has been a major theme of several NCVHS Roundtables on community health data held between 2011 and 2016.[18] To meet the growing need for detailed local data, community groups and agencies are increasingly collecting their own primary data or finding and creatively repurposing existing data to augment secondary sources.

NCVHS has observed that "growing linkages and granularity can - and should - heighten privacy concerns" when there is a risk of identifying individuals, stigmatizing groups, or otherwise compromising privacy.[19] This perspective must be kept in mind when considering the data needed to tackle community health concerns such as childhood obesity, a topic explored in the scenarios and the use case presented in Appendix 2.

[17] http://www.ncvhs.hhs.gov/111213chip.pdf (p. 23)

[18] http://www.ncvhs.hhs.gov/130430sm.pdf & 2014 http://www.ncvhs.hhs.gov/supporting-community-data-engagement-an-ncvhs-roundtable/

[19] http://www.ncvhs.hhs.gov/wp-content/uploads/2013/12/Toolkit-for-Communities.pdf (p. 28)

The goal of providing a taxonomy of methods for moving among different levels of granularity is to make it possible to work with data at multiple levels of aggregation while taking into account relevant social constructs and constraints. The need for granular data is a function of the specific uses of the data, such as evaluation, research or intervention. Other factors with influence on how granular the data need to be include the nature of the data source, methods, requirements, whether the focus is an individual or an institution, and social structures. Stewardship responsibilities also vary across these dimensions and others, with differences in sensitivity regarding privacy and tolerance for disclosure.

The differences between dense urban and sparse rural populations have an impact on the appropriate data collection infrastructure and consent mechanism, as well as on the risk of harm. For example, different data stewardship techniques are required when working with data from a small group of 100 individuals with a rare disease drawn from a large, geographically dispersed population of 10 million, than when working with data from the same size group of individuals drawn from a rural county due to privacy concerns.

Some of the relevant granularity variables are shown in Exhibit 6 which explores granularity in terms of contrasts between analyzing a sub-population with a rare condition and analyzing any kind of geographic community.

Exhibit 6. Granularity Variables

Variables	Analysis of geographic communities of any size or density	Analysis of a sub-population with a rare condition
Data type	Passive collection Environmental Socioeconomic & cultural	High individual density of data Environmental
Geography	More relevant	Less relevant
Infrastructure	Common	Specialized
Analysis	Large: New methods, new data types Population intervention	Statistical methods for small groups "Classic" analysis Individual intervention
Stewardship	Population/political accountability	Higher risk of exposure, but maybe also need-based tolerance

4. Paths Forward and Vision of the End Game:

With this white paper, NCVHS seeks to inspire the Federal Government and the data supplier ecosystem *in toto* to elaborate and promulgate the Health Data Framework.

It is widely recognized that the lack of interoperability is a major obstacle to the convergence critical for achieving the Triple Aim of improved patient experience, improved population health, and reduced per capita cost of healthcare.[20] Some of the critical types of data interoperability that can be enhanced by the Framework are shown in Exhibit 7. The interplay between the Framework's Data Structure and Methods Taxonomy propose a path toward enhancing interoperability.

Exhibit 7. Types of Data Interoperability

- **Syntactic**: Linking industry-adopted standards formally recognized by a standard-making body to the data set being collected/exchanged (e.g., version of message format or content standard).
- **Semantic:** Synchronizing definitions of concepts, terms, and variables (e.g., defining smoking or functional status).
- **Privacy**: Aligning health information privacy policies across health and information systems to allow the collection, use, and disclosure of information (e.g., matching primary data source restrictions to threshold for secondary use).
- **Security**: Using comparable health information security policies and practices across systems to ensure consistent availability, confidentiality, and integrity of health information (e.g., specific security rule)
- **Granularity:** Coordinating units of geography for which data are available (e.g., individual through national).
- **Time**: Aligning currency of data and periodicity of data collection (e.g., real time data and how often collected).
- **Content domain**: Aligning areas of focus (e.g., clinical indicators, risk behaviors, social/economic context, environmental factors, community assets).
- **Analytic interoperability**: Aligning tools for data manipulation (e.g., GIS, simple statistical software, WDQS).

The Data Structure and the Methods Taxonomy proposed here can be used to begin, as a source of metadata, to make more explicit the scope and characteristics of data sets. At first glance, use of the Health Data Framework in this way may seem daunting. To the contrary, the data set developer knows its measurement scope, purpose, whether it contains self-reports by individuals or responses from administrative staff, etc. They can skip non-relevant subcategories — if the set is heterogeneous they can pick as many as apply.

[20] Interoperability is defined here as the ability of all of the actors who work to improve the health of individuals and populations (from the community to the international level), including patients and other lay people, and of different information systems and applications, to communicate, exchange data, and use the information that has been exchanged.

These draft resources are designed to be extensible. New categories can be added to a level, or a category can be subdivided by adding an additional level. Accordingly, they can be elaborated through centralized, consortia, or open-source approaches.

Consider the following scenario of "the end game": It is January 2019. The Framework has been fully developed. The National Library of Medicine hosts the Data Structure and Methods Taxonomy knowledge sources (as they host the Unified Medical Language System (UMLS) meta-thesaurus and related resources). Public and private data providers tag data sets to the appropriate coordinates along the dimensions of the Data Structure and to characteristics in the Methods Taxonomy. A rich ecosystem of commercial companies and consortia develop analytic workbenches built on top of these classification resources.

A community coalition[21] has targeted childhood obesity as its top priority after talking with community members, and analyzing data on health disparities and assets. As they explore national survey data from their county, they zoom down the geographic dimension to a dataset aggregated by the State Health Department at the census tract level to identify a hotspot (high incidence and prevalence of childhood obesity) and the nearest coldspot (low incidence and prevalence). Then they zoom out to the neighborhood level to a dataset showing what community organizations these two areas have in common. Next they move from the outcomes level on the population health measurement dimension to the intervention level. They see a primary data set with school programs whose purpose is consistent with the purpose of the coalition's secondary use in reducing childhood obesity. Accordingly, the analytic workbench grants access to the data set, unpacks the data with the correct archival version of content standard, and creates a visualization appropriate to the data set and the secondary use.

[21] This scenario is based on the Use Case in Appendix 2

Appendix 1. Methods Taxonomy (v.1.3, 4/17/2015) [22]

1. Data source characteristics			
a. Type of Data Source	i. Electronic health records	1) Care provider	
		2) Health Information Exchange	a) Regional b) Laboratory reporting c) E-prescribing d) Etc.
		3) Etc.	
	ii. Personal journal or health record	1) TBD	
	iii. Domain-specific measurement instruments	1) TBD	
	iv. Community data sets	1) TBD	
	v. National surveys	1) TBD	
	vi. Payor Datasets	1) Medicare 2) Medicaid 3) Etc.	
	vii. Social network data sets	1) TBD	
	viii. Economic actor data set		
	ix. Etc		
b. Original collector and aggregator	i. Government	1) Jurisdiction	a) Federal b) State c) Sub state
		2) Type of authority	a) Public health authority b) Non-public health agencies (e.g. social c) Law Enforcement d) Environmental authority
	ii. Health plan		
	iii. Health care provider		
	iv. Individual member of the public		
	v. Economic actors - corporate and private		
	vi. Etc.		
c. Purpose of collection	i. TBD		
d. Method of collection	i. TBD		
e. Voice	i. Sellf-report		
	ii. Administrative staff		
	iii. Trained observer		
	iv. Passive collection (devices)		
	v. Etc.		
f. Granularity	i. Collection level		
	ii. Aggregtion level		
	iii. Minimum # of individuals represented in a sample		
g. Primary users	i. TBD		
h. Primary uses	i. TBD		
i. Applicable regulations	i. Protected health information (HIPPA privacy rule)		
	ii. Electronic identifiable health information (HIPPA		
	iii. Family educational rights and privacy act (FERPA)		
	iv. State regulations		
	v. Institutional review board (IRB)		
	vi. Etc.		
j. Identification status	i. Individually - identifiable data		
	ii. De-identified data (HIPPA definition)		
	iii. Anonymized date	1) No linkage possible 2) Re-linkable data 3) Linked with protected key	
k. Consent provided at the time of data collection	i. No consent by the individual		
	ii. Consent by the individual	1) Broad and unspecified 2) Time-limited consent 3) Consented for partial, source specific 4) Consented for the particular type of use	

[22] Limited to 4 levels of subcategories

1. Data source characteristics, continued			
l. Applicable Standards	i. Content	1) TBD	
	ii. Messaging	1) TBD	
	iii. Etc.		
m. Demographic representation	i. Age		
	ii. Race		
	iii. Gender		
	iv. SES		
	v. Insurance status		
	vi. Etc.		
n. Vulnerable populations included	i. Prisoners		
	ii. Pregnant women		
	iii. Undocumented immigrants		
	iv. Etc.		
o. Population health measures	i. *[Link to appropriate levels in DS population health*		
	ii. *[Link to appropriate levels in DS determinants of*		
	iii. *[Link to apprtopriate levels in DS pathophysiology*		
p. Timing	i. Cross-sectional		
	ii. Longitudinal		
q. Accuracy	i. Level of confidence		
	ii. Etc.		
r. Completeness	i. TBD		
s. Timeliness	i. Rate of chage		
	ii. Shelf life		
	iii. Acuity of need		
	iv. Lag time		
	v. Background rate of change		
t. Limitations	i. Provisional vs. preliminary		
	ii. Open vs. closed		
	iii. Etc.		
u. Biases	i. TBD		
2. Secondary data use			
a. Purpose	i. TBD		
b. Users	i. [*Re-use categories under original collector or*		
c. Uses	i. Healthcare		
	ii. Public health		
	iii. Social services	1) Abuse, neglect or domestic violence	a) Child abuse or neglect
		2) Workplace safety	
	iv. Law Enforcement		
	v. Etc.		
d. Granularity	i. TBD		
e. Timing	i. TBD		
f. Timeliness	i. TBD		
g. Etc.	i. TBD		
3. Analytic and Visualization Methods			
a. Data collection	i. Qualitative	1) Key informant interviews	
		2) Opinion surveys	
		3) Focus groups	
	ii. Quantitative	1) Sample survey data collection	a) Sample Design
			b) Sampling issues
		2) Causal statistical studies	a) Experimental
			b) Observational
	iii. Granularity		
b. Analysis	i. Descriptive presentation	1) Counts	
		2) Trends	
		3) Periodicity	
		4) Rates/proportions/percentages	
	ii. Statistics	1) Univariate analysis	
		2) Bivariate analysis	
		3) Multivariate	

3. Analytic and Visualization Methods, continued			
b. Analysis, continued	iii. Inferential statistics	1) Modeling assumptions	a) Fully parametric b) Non-parametric c) Semi-parametric
		2) Inference methods	a) Classical b) Bayesian c) Other
	iv. Statistical classification & machine learning	1) Linear classifiers	a) Logistic regression b) Naïve Bayes classifier c) Etc.
		2) Support vector machines	
		3) Quadratic classifiers	
		4) Decision trees	
		5) Neural networks	
		6) Etc.	
	v. Simulation	1) Discrete event	
		2) Queuing networks	
		3) Etc.	
c. Visualization	i. Graphs		
	ii. Maps		
	iii. Multi-dimensional plots		
	iv. Nodes and links		
	v. Trees		
	vi. Etc.		

4. Data stewardship principles		
a. Openess, transparency and choice	i. Policies and practices regarding community and personal data are publicly available	1) Effective channels for communicating with community stakeholders are 2) Data subjects can learn about what uses are being made of data and how the data are being protected 3) If appropriate, data subjects are informed of results of data use
	ii. Data are obtained through legal means	1) Data stewards have a process in place for reviewing the legality of proposed data gathering and use 2) Data stewards have a process in place for detecting malfeasance and taking action if any occurs. This includes malfeasance by third parties to whom data have been transferred (e.g. violation of a DUA)
	iii. Communities that are subject of data use are provided notice	1) Notice processes are in place 2) Q & A processes are in place
	iv. Individual whose personal health data are to be used has right to consent.	1) Notice of privacy practices is available 2) Consent processes are in place
	v. Individuals have the right to opt in or opt out of community data use projects	1) Notice processes are in place 2) Default approach is explicit
b. Purpose specification	i. Data users engage community stakeholders to define purpose	1) Processes for dialogue with stakeholders about purpose 2) Stakeholders understand that certain data are required by law
	ii. Data sources and types are fit for the purpose	1) Understand fit for use limitations in available data sets
	iii. There is clarity in plans to repurpose data or to use repurposed data	1) Consider stakeholder concerns when repurposing data 2) Understand legal restrictions on repurposing data

4. Data stewardship principles, continued		
c. Colletions and use limitation	i. The data collected are limited to what is needed for the intended use	
d. Data quality	i. Processes for assessing the quality characteristics of data sets to support intended use	1) Assess data for accuracy 2) Determine whether data are valid and reliable 3) Data are timely and complete
	ii. Processes for collecting and preparing data for use	1) Processes to assess trustworthiness of data sources 2) Processes for merging data sets 3) Processes for cleansing data
	iii. Processes for effective analysis and use of data	
e. Security safeguards	i. Appropriate use of de-identified data	
	ii. Avoidance of re-identification	1) Data stewards have in place practices to assess whether re-identification is occurring 2) Data stewards have process for detecting unauthorized re-identification and who is responsible for it
	iii. Encryption protections password	
	iv. Training	
	v. Storage	
f. Accountability	i. Responsibility is assigned for each phase of data lifecycle	1) Clarity of assignment to responsible entities or individuals 2) Consequences of accountability failure are delineated
	ii. Data use agreements (DUAs) are used when appropriate	1) Knowledge of laws and regulations regarding data sharing 2) Understanding of DUA provisions 3) Ability to assess DUA compliance 4) Methods for dealing with non-compliance
	iii. Forms of agreement other than DUAs are used as appropriate	1) Provisions of agreement are understood by all parties 2) There are mechanisms to assess compliance
5. Standards		
a. TBD		
6. Additional categories		
a. TBD		

APPENDIX 2. Use Case

A Community Childhood Obesity Reduction Project

This Use Case illustrates the complexities associated with addressing multiple determinants, capturing a range of perspectives on community health, and dealing with the variations in data sources and availability.

In this scenario, which is also explored elsewhere in this paper, a community has targeted childhood obesity reduction as its top priority after conducting an assessment process, talking with community members, and analyzing the data on health, disparities, and assets. The use Case incorporates the perspectives of four groups of coalition members (represented in the columns), each of which encompasses a constituency and set of actors, an institution or sector, an area of expertise and responsibility, and/or a set of information assets. Using illustrative rather than exhaustive lists, the table below identifies some of the information such a community would need to guide decision-making and achieve the stated goals. It divides the data into three categories, all of which the Data Structure comprises: (1) data that community agents *already possess* from existing internal and external sources; (2) other data that they are *able to obtain*; and (3) *needed* data that must be found from new and potentially unusual sources.

Perspectives ➔	Community	Schools	Health Care Teams	Public Health
Goals	▪ Community culture of wellness ▪ Shift BMI distribution in targeted age groups in 5 years, e.g., Fewer overweight kids entering kindergarten ▪ Increase activity level in families with young children in 5 years ▪ Change in attitudes toward diet & exercise in kids entering HS ▪ Healthier kids	▪ Kids that are "fit to learn" ▪ Shift BMI distribution of kids in the district ▪ Increase activity of kids in district ▪ Increase in healthy lunches	▪ Reduce incidence of obesity related co-morbidities for patients in their care ▪ Stable BMI appropriate to body frame ▪ Trusted point of access to health care for all members of families in their care ▪ Increase awareness of clinical team about community resources	▪ Decrease disparities in nutrition, activity & obesity ▪ Decrease morbidity & mortality ▪ Increase community awareness of obesity risk & trends ▪ Increase community awareness of barriers to proper nutrition & activity
Program Design Questions				
Target population(s)?	▪ Pre-kindergarten ▪ Schools ▪ Day care ▪ Churches ▪ Malls	▪ Kindergarten ▪ Lower elementary ▪ Upper elementary ▪ High school	▪ Perinatal families ▪ Pediatric age groups ▪ Adolescent patients	▪ State ▪ County ▪ Neighborhoods
Intervention(s)?	▪ ID most effective "upstream" interventions ▪ Public awareness campaigns - wellness is "hip", target at young parents and kids; obesity risk ▪ Ban ads for unhealthy foods targeting youth ▪ Programs for key life transitions, birth, entry into kindergarten, elementary, health care ▪ Day care nutrition guidelines ▪ Fresh food markets near school ▪ Healthy supper clubs in churches, community centers, grocery stores ▪ Clean up, light & monitor parks & playgrounds ▪ Access to safe recreation areas for kids & families ▪ 1k steps/day campaign with pedometers ▪ Provide web-based referral to community wellness resources for public, school nurses, HCPs	▪ ID most effective school based interventions ▪ Affordability in healthy lunches ▪ Healthy snack machines ▪ Student/parent healthy supper classes ▪ Physical education interventions and activities ▪ Increase activity in extracurricular activities ▪ School based wellness coordinators & nurses	▪ ID most effective practice-based interventions ▪ Training & information for clinicians, technical resources & incentives ▪ Screen for diet and activity ▪ Include nutritional & activity coaching in assessment of developmental milestones ▪ Referrals to community resources for wellness & life change ▪ Know who adolescent patients are; demonstrate sustained long relationship with them; review their "journals" of relevant data	▪ ID most effective public health interventions ▪ Disseminate data on prevalence & risks of childhood obesity ▪ Review literature on what works & provide good information for all coalition members ▪ Promote awareness & convene stakeholders to share perspectives ▪ Improve parent awareness of obesity risks ▪ Establish trust as data steward for community, collect missing data & convene discussion of meaning of data

Process and outcome measures?	▪ O: Weight, BMI by age cohort & neighborhood ▪ P: media appearances; O: awareness of risks, attitudes ▪ P: minsters pitched to; O: churches adopting interventions ▪ P: # gardens planted ▪ O: time parks and playgrounds available, # kids participating ▪ O: pre-post survey family perceptions of change	▪ P: % of schools in district participating ▪ P: % school measurement of BMI at start of yr, O:% elevated BMI entering next grade ▪ P: # students participating by type of intervention ▪ O:% healthy lunches ▪ O:Minutes of in-school activity trends by school & age cohort ▪ O:BMI trends for studies by school & age cohort	▪ P: % screened, % coached, % referred to community resources ▪ O: Distribution of # days w active exercise & minutes/day in patients under care ▪ O:Distribution of fruit and vegetable consumption in patients under care ▪ O:Distribution of BMI trends in families under care ▪ O: Diabetes-2 prevalence in families under care for 2 yrs, 5 yrs, 10 yrs	▪ P: % of effective interventions implemented ▪ O: Weight, BMI by age cohort & neighborhood
Relevant data they have (& may supply to partners)	▪ Bike trails, walking, other rec opportunities & spaces, safe & clean	▪ Location, resources, staff, programs ▪ Catchment area served	▪ BMI, ht, wt, BP for patients under care ▪ Payer	▪ Incidence and prevalence of obesity by census tract ▪ # & types of providers avail ▪ # households ▪ WIC kids' BMI ▪ SES data on community, demographics, housing stock, density, public safety ▪ Unemployment, Medicaid, free lunch, etc.
Additional primary data they can obtain	▪ Snack food revenues ▪ Neighborhood assessment of kid activity, qualitative data, # gunshots, space avail.	▪ BMI ▪ Belly circumference ▪ Hours of exercise/day ▪ Calories of school meals ▪ # Vending machines SES factors	▪ Complications of childhood obesity ▪ Activity & nutrition screening ▪ More thorough information on family, including history of diabetes, culture/attitudes	▪ Survey available space for recreation & current activity there ▪ % w/in 1 mi of walking ▪ School, healthy food
Needed data outside usual sources	▪ Retail ▪ Church health fairs ▪ Salons ▪ Malls ▪ Barber shops	▪ School-based BMI measures ▪ Children at risk for obesity	▪ Comparison data on children in their practice vs others; ▪ Trend data on childhood obesity	▪ % households with trusted place of entry into the health system ▪ SES predictors ▪ Accurate information on community resources to deal with childhood obesity; ▪ Social media sources, every kid ▪ Need for detailed, comparative data across neighborhoods to show disparities