Toolkit for Communities Using Health Data

How to collect, use, protect, and share data responsibly

A Report from the
National Committee on Vital and Health Statistics

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
Centers for Disease Control and Prevention
National Center for Health Statistics

Toolkit for Communities Using Health Data

How to collect, use, protect, and share data responsibly

National Committee on Vital Statistics
May 2015

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
Centers for Disease Control and Prevention
National Center for Health Statistics

Table of Contents

Introduction

The National Committee on Vital and Health Statistics (NCVHS) is the U.S. Department of Health and Human Services' (HHS) statutory public advisory body on health data, statistics, and national health information policy. NCVHS has historically made recommendations regarding stewardship of health information collection, use, and disclosure.

In recent years, NCVHS hearings and roundtable discussions about how communities are using data to improve health at the individual, subgroup, and community levels have shown the need for guidance on the meaning and application of data stewardship. These efforts have focused on the needs of community-level organizations. NCVHS created the Community Data User Toolkit to be a substantive introduction to the elements of data stewardship for communities that want to use data.

In this document, a community is defined broadly as a formal or informal group with a shared interest, which could be defined by a shared characteristic such as geography, race or ethnicity, a shared medical diagnosis, or a combination of characteristics. For example, a community could be a neighborhood in a city, an online community of individuals affected by cancer, or a racial subgroup within a city.

This document also uses the term data broadly. Communities may use many different types and sources of data to promote the health of the community, subgroups, or individuals. Some data will be related to health conditions, but other data could relate to environmental factors, such as locations of grocery stores or access to safe walking routes. Data related to health conditions could come to the community as aggregated data collected for other purposes, such as disease surveillance. Other health data could be abstracted from patient medical records, or collected by the community user through a survey or some other mechanism.

Community groups today are using data to tackle important health issues in ways that were not even imagined a few years ago. In the past, access was largely limited to government-based public health agencies or health care systems. Now communities can access data because data availability has exploded, particularly data in digital formats. Federal and state governments, local health information exchanges, and other organizations have data that could be made available to promote community and individual health. If used effectively,

"Community groups today are using data to tackle important health issues in ways that were not even imagined a few years ago."

data may help improve communities' understanding of:

- Health of the community and members of the community
- Health challenges facing the community
- Health promotion successes within the community
- Opportunities to improve the health of the community as a whole and the health of individuals living in the community

Many organizations have data that may be available for communities to use. These organizations may also provide tools and guidance for communities wanting to use their data. This Toolkit includes important themes in stewardship—proper data protection and use—and, where relevant, refers community data users to some of these resources.

"Failure to use good stewardship practices could harm individuals or communities."

Effective data use requires effective stewardship practices. Failure to use good stewardship practices could harm individuals or communities. Improper data handling or the failure to protect individuals' privacy or confidentiality could limit participation and impede the use of data.

This Toolkit was created to support communities that are using data by promoting sound stewardship practices, while helping them avoid the missteps and potential harm that can result when data users do not follow sound data stewardship practices. The Toolkit is not meant to provide a comprehensive explanation of every aspect of data stewardship, nor is it meant to be a substitute for legal counsel or expertise in data collection, use, disclosure, or security. We hope that communities will find this Toolkit helpful as they continue to use data to improve health.

Why a Toolkit and Why Now?

Technology is changing everything. Thanks to technology, information is now developed, shared, and used in new ways. Communities have opportunities to use data to improve community health and the health of individuals living in the community, opportunities that did not exist in the past.

Another less obvious opportunity comes from the growing realization that communities are in the best position to identify the challenges they face and the strengths they enjoy. Therefore, communities themselves may be best positioned to find the most effective ways to use data to understand and address their health needs.

By bringing technology and community-defined concerns together, data can now be effectively used to address community-defined problems and to secure and protect

community assets. Measurement and analysis are necessary (not optional) pieces of the puzzle that allow communities to know where, and why, health is improving or declining. In addition to addressing what is known, data have the potential to allow communities to discover unknown factors that matter to them. Data also have the potential to yield findings that may be surprising to, or unwelcomed by, community members.

Done right, using data builds the trust that is essential for finding, defining, exploring, strengthening, and improving health at the community and individual levels.

What the Toolkit Does

The Toolkit briefly introduces each important principle of data stewardship for communities using health data.[1] It provides both broad background information and tips for data users. Descriptions of stewardship principles are provided, along with checklists for each principle.

As experienced data stewards know, and as emerging data stewards will learn, the different principles described in the Toolkit do not divide neatly into separate categories, but rather overlap and intertwine. For example, the two principles Openness, Transparency, and Choice and Community and Individual Engagement and Participation, are relevant across every step in the stewardship framework and throughout the data lifecycle. To the extent that principles are interrelated, they are introduced in a unique section, but are also referenced in sections addressing other topics when relevant.

Different types of data trigger different approaches to stewardship, with the burdens of stewardship and the balancing of interests changing from one type of data to another. Because of its likely sensitive character, health information presents important issues for data stewards. A data steward investigating the density of grocery stores in a neighborhood is not likely to encounter major concerns about privacy or confidentiality. But a data steward who wants to use personally identifiable health records that contain the results of genetic testing is very likely to encounter those concerns. The primary focus of the Toolkit is health data, which will typically require rigorous attention to all of the elements of data stewardship. However, the principles in the Toolkit may be more broadly applicable to many different types of data and their uses for communities.

"Measurement and analysis are necessary (not optional) pieces of the puzzle that allow communities to know where, and why, health is improving or declining."

1 For a more detailed discussion of the NCVHS framework of stewardship principles, see National Committee on Vital and Health Statistics, Letter to Secretary Kathleen Sebelius, "A Stewardship Framework for the Use of Community Health Data," (Dec. 5, 2012) available from: http://www.ncvhs.hhs.gov/wp-content/uploads/2014/05/121205lt.pdf.

Appendices

Appendices are provided with supplemental information, including:

- Definitions
- Legal Considerations
- Case Studies
- Checklists

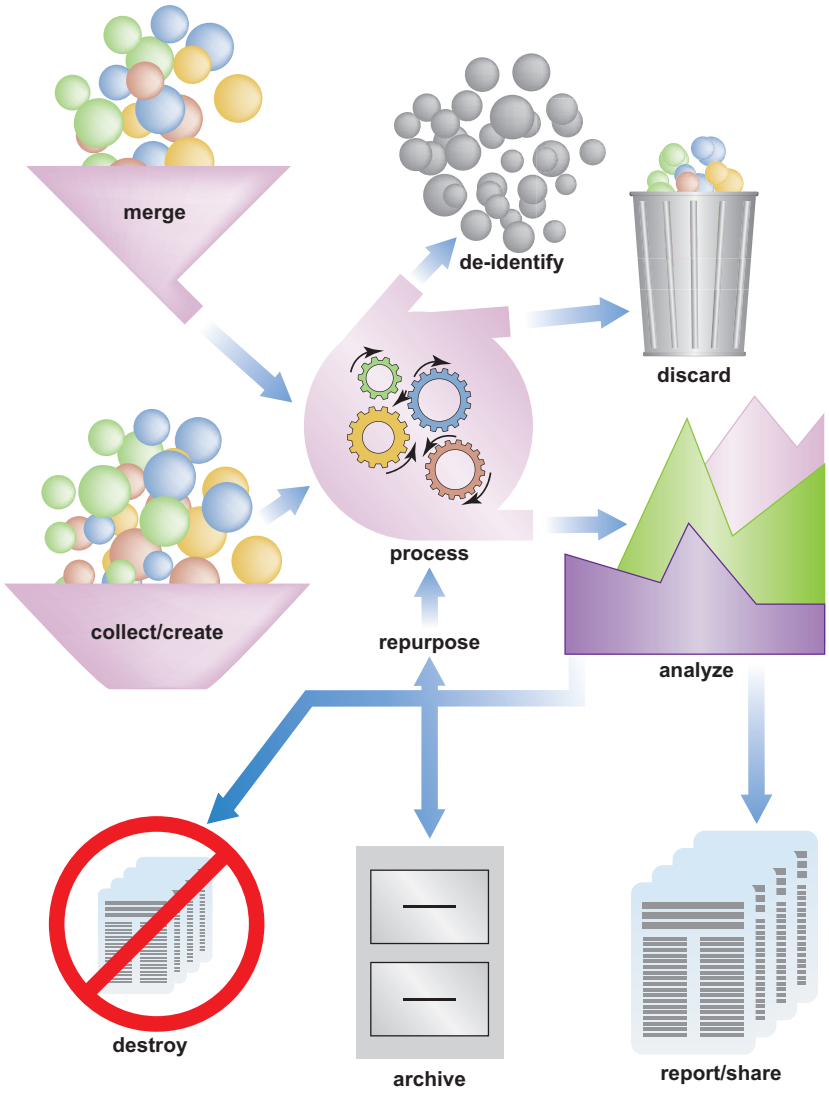
Data Lifecycle

Data have a lifecycle, represented in the figure below. Effective stewardship extends to all lifecycle phases. Examples of communities using data across the lifecycle are provided throughout the Toolkit.

Not all data move through all parts of the lifecycle. Some are collected and never analyzed. Some analysis fails to produce reportable results. Some data are never destroyed but are stored in perpetuity.

There are also steps that communities using data to improve health must take that are outside of the data lifecycle, such as doing a literature review to learn about the current knowledge on the topic and to better frame the purpose of the inquiry.

Lifecycle of Data: From Collection to Disposition



"Not all data move through all parts of the lifecycle. Some are collected and never analyzed. Some analysis fails to produce reportable results. Some data are never destroyed but are stored in perpetuity."

Original or Repurposed Data

Community health data can be either original or repurposed.

Original data are gathered for an initially specified purpose; they are data that did not previously exist. For example, original data may be collected through a survey of community members about access to fresh fruits and vegetables in local markets, observation of activities of children in a playground, or new survey research on the incidence of a health problem in the community.

"Uses of repurposed health data have expanded sharply with access to digital data from electronic health records and other information technology; these uses are likely to continue to expand."

Repurposed data are collected for one purpose and then used for a different purpose. Communities may want to repurpose data from a variety of sources.

Until recently, the data in patient medical records were used primarily for patient care, payment, and health care institution operations. Data abstracted from paper medical records were used for research and other purposes, but it was costly and difficult to extract data. Uses of repurposed health data have expanded sharply with access to digital data from electronic health records and other information technology; these uses are likely to continue to expand.

For example, an individual may complete a questionnaire about health status as part of a doctor's visit that is then entered into the history and physical section of the electronic medical record. Later, relevant responses are pulled from the electronic health records of all patients who completed the questionnaire into a new data set that will be used to evaluate the prevalence of a condition among community members. The responses to the initial health questionnaire collected for the purpose of treatment are repurposed to determine disease prevalence.

Communities also often repurpose public health data generated by local, state, and federal government agencies. For example, communities might investigate changes in teen birth rates, opiate deaths, cancer clusters, or suicide rates. In doing so, they might use data that were collected for one purpose, such as to determine cause of death, for another purpose, such as to explore correlations between social factors and suicide. They might also combine these public health data sets with other available data or data they collect themselves.

Relationship Between Technology and the Data Lifecycle

Information technology has greatly changed how data are managed at all lifecycle stages from creation to destruction or archive. Technology speeds the capture of data and makes it available for use sooner. It can help to keep a description of the characteristics of data, called metadata, including who collected the data, when the data were collected, what permissions or restrictions are attached to the data, flaws or limitations of the data, and other such characteristics. Technology can also be used to set up rules for data capture and collection, processing, storage, exchange, and dissemination in ways not imagined just a few years ago.

New technology enables users to:

- Store large amounts of electronic data
- Process and analyze large data sets efficiently
- Enrich data sets by merging data from different sources
- Monitor trends over time to track changes
- Repurpose data in ways not conceived when the data were collected
- Access data remotely
- Copy or transmit data rapidly

Electronic health records are, like paper medical records, used initially to support the delivery of patient care, payment, provider operations, and quality improvement, but the electronic format makes the records more useful to researchers, public health agencies, and communities seeking to improve the health of individuals and communities. For example, electronic claims data are increasingly used to track public health issues and to allocate limited funds to areas of greatest potential impact.

Technological advances offer both opportunities and risks to communities using health data.

"Technological advances offer both opportunities and risks to communities using health data."

"The Toolkit can help data users take advantage of the opportunities that technology offers while avoiding risks."

Opportunities include:

- Understanding health at a more detailed or granular level, such as geo mapping health data to show how disease affects individuals living on a specific block within a community
- Evaluating the impact of programs on health by linking data about who received an intervention with data from a communitywide health information exchange or from repurposed claims data

Risks include:

- Data breaches: Data security is challenging, even for large companies and governments with substantial resources.
- Data elements: They can appear to be the same but have different meanings across systems, causing incorrect interpretation.
- Repurposing: This can cause harm when it happens without appropriately engaging and involving individuals and communities, as shown in many of the Case Studies described later in this Toolkit (Appendix C).
- Problematic inferences due to analysis of electronically processed data: These may result in social stigma and harm to the reputations of wrongly categorized individuals.

The Toolkit can help data users take advantage of the opportunities that technology offers while avoiding risks.

Governmental and Nongovernmental Data Collectors and Users

Data stewardship for nongovernmental data collectors or users has much in common with, but is not the same as, data stewardship for governmental data collectors or users. Still, both government and nongovernment data stewards must follow the laws, regulations, and policies designed to protect the privacy and confidentiality of individuals and the integrity and security of the data. Governmental data stewards hold data in trust for the public; they have an affirmative duty to serve the public by openly and transparently sharing data. Nongovernmental data users and collectors do not share that affirmative duty, although sharing data to serve the community and public good is consistent with stewardship principles.

Data Stewardship

Data stewardship is a responsibility, guided by principles and practices, to ensure the knowledgeable and appropriate use of data. More specifically, stewardship of health data recognizes the benefits to society of using personal health information to improve understanding of health and health care, while at the same time respecting individuals' privacy and confidentiality. The individual elements of data stewardship are driven by ethical imperatives that require data users to respect the individuals who are the subjects of health data.

Many people touch data as it moves through its life cycle, and each person who touches the data should have an awareness of relevant stewardship principles and practices.

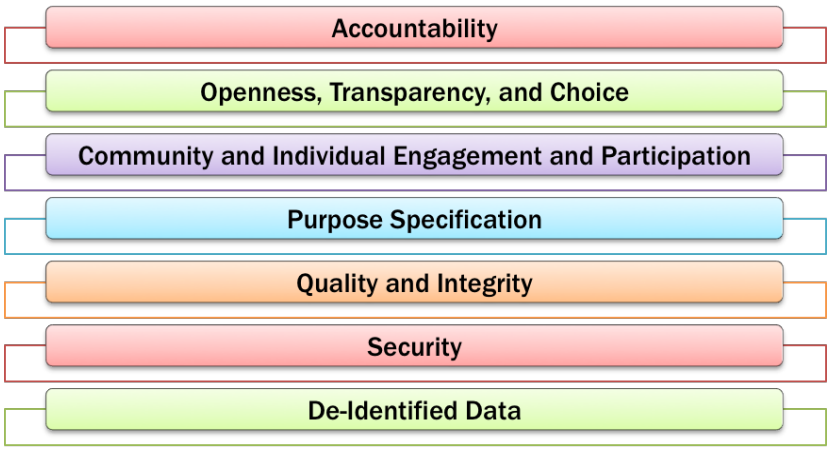
Communities are encouraged to use data to improve health, while following responsible data use practices so that individuals or groups whose data are used by communities to improve health can trust that private or confidential information is being used appropriately.

Nonlinear, Overlapping Concepts

The figure showing the elements of data stewardship below suggests that stewardship elements follow a certain order. In reality, as noted throughout the Toolkit, elements overlap, and the stewardship process may require data users to loop back or jump forward as needed.

"The individual elements of data stewardship are driven by ethical imperatives that require data users to respect the individuals who are the subjects of health data."

Principles of Data Stewardship



Accountability

The first thing a community should do when thinking about a new data analysis project is assign responsibility for accountability for all parts of the project. Accountability means that an individual or entity is responsible for:

- Ensuring appropriate collection or creation, use, disclosure, and retention of data through policies and practices, and
- Establishing mechanisms to find and respond to any failure to follow policy and procedures.

It should be made clear who is accountable at each phase of the data lifecycle—from project planning, through initial collection and use, to data destruction, storage, or repurposing. Different people or entities might be accountable for different phases, but this should be made explicit. Accountability for each aspect of data stewardship should also be clearly assigned so data users understand who is responsible. If there is a failure of accountability, the responsible individual or entity should face appropriate consequences and provide remediation to individuals affected by the lapse.

"Failure to identify and address concerns regarding proper data stewardship may lead to downstream consequences, some mild, others quite serious."

Failure to identify and address concerns regarding proper data stewardship may lead to downstream consequences, some mild, others quite serious.

Data Use Agreements and Accountability

Data use agreements (DUAs) can help an entity enforce the various privileges and obligations involved in sharing or obtaining data. In combination with other protective measures, these agreements can be useful tools for managing accountability.

DUAs are not a guarantee that data will not be misused. With or without statutory authority, an entity that shares data may need to take legal steps to enforce a DUA if a data user violates the agreement.

Considerations in Signing DUAs

A DUA is a contract—a legal document with legal implications. It should not be taken lightly. If a data user is asked to sign a DUA, the user should consider the accountability checklist items outlined in Appendix D. An organization that is asked to sign a DUA should understand what the DUA requires of it and should be confident that it can meet those requirements. If an organization has questions or concerns about the document, it may be useful to consult legal counsel.

Summary

- Accountability may lie with an individual or entity.
- Different people may be accountable for different phases of the data lifecycle or different stewardship elements.
- An accountable individual or entity should be named and held responsible for stewardship.
- DUAs are one way to establish accountability ground rules among data users.

Accountability ombudsman

Vanderbilt University, a member of the Electronic Medical Records and Genomics (eMERGE) Network, identified accountable individuals or groups for each stage in the data lifecycle, but found that this was not enough. Communities that Vanderbilt worked with needed one person who could be their accountability contact. The eMERGE network at Vanderbilt appointed an individual to explain the organization's accountability policies and procedures to people in the community and who could ensure that their concerns would reach the accountable person. Members of the eMERGE Network describe this approach as "a lifesaver."

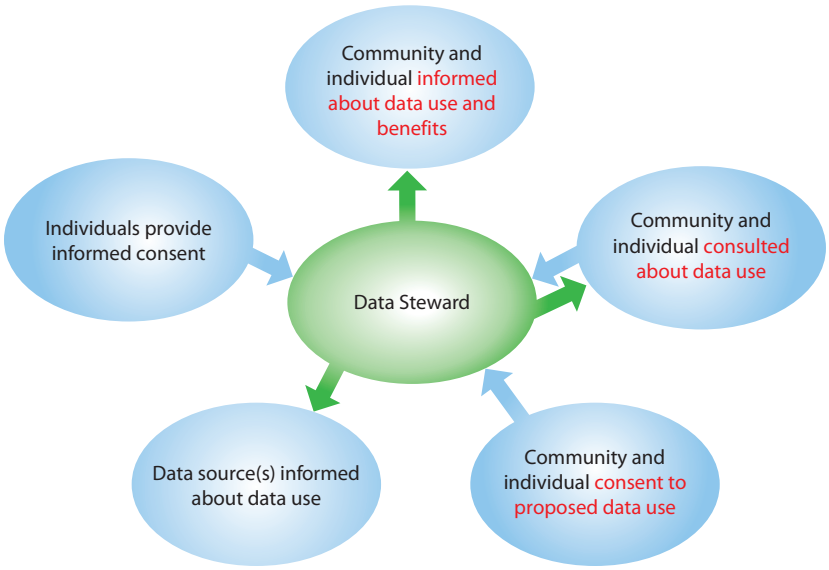
Openness, Transparency, and Choice

Openness, transparency, and choice promote trust among data users, data sources, individuals, and communities. If data users are not open and transparent or if they do not offer choices to individuals and communities when required or appropriate, this can create unwelcome surprises, destroy trust, and may even reduce the ability to use health data to improve health in the future. The Toolkit includes examples of such failures as cautionary case studies.

Community engagement supports openness, transparency, and choice. For example, community leaders, neighbors, or advisory boards can serve as conduits for notice to community members. Communities can also provide information to data users about how community members view the data use, the level of disclosure, and the range of choices necessary to maintain the community's trust, as depicted in the following diagram.

"Community engagement alone may not, however, be enough to ensure openness, transparency, and choice in cases where individuals' preferences are not the same as the interests of the community."

Advancing Openness, Transparency, and Choice



Community engagement alone may not, however, be enough to ensure openness, transparency, and choice in cases where individuals' preferences are not the same as the interests of the community. To maintain trust, data users must be open about expectations of data use.

Notice and consent are at the heart of openness, transparency, and choice.

Notice is information provided to the community about data use.

Consent is the process of getting permission from a community or individual to use data.

Notice

Data users should provide individuals and communities with notice about:

- What information is being collected
- Goals and potential benefits of data use
- Risks of data use

Communities and individuals whose data will be used should be able to ask questions about, comment on, or object to data use. Data users may also need to give sources of data, such as health care providers, public health agencies, or researchers, the same type of information.

"Communities and individuals whose data will be used should be able to ask questions about, comment on, or object to data use."

Individual notice

Individual notice may be needed when those whose data are being used are identifiable, for example, by name or home address, and when the risk of compromising privacy or confidentiality or stigmatizing an individual or small group is high.

Direct Individual Notice

If data users plan to use protected, personally identifiable data without other prior notice, they may need to provide individual notice. In some instances, laws or regulations require individual notice, but stewardship practices also may warrant individual notice if the risk of violating an individual's confidentiality or privacy is significant, or if disclosure could cause harm. Data users may provide individual notice through a telephone call, a face-to-face encounter, e-mail, or traditional mail. Mail is the most costly and burdensome form of notice. For example, a data user may have a name but no address, so the data user would spend time and resources finding the person's address or other means of contact. Even where addresses or telephone numbers are available, it is costly to place phone calls or to mail notifications to individuals for more than a small number of individuals.

Engaging the community to determine type of notice

MyHealth Access, a non-profit health information exchange in Oklahoma, took on the challenge of engaging the residents of Tulsa. The organization's Privacy and Security Committee explored two distinct choices: notice through the newspaper or personal notification. They conducted focus groups in doctors' waiting rooms, asking, "Where do you want to learn about the sharing of your data?" Patients did not want to read about it in the newspaper for a number of reasons. Rather, they wanted to receive notice about data use in the doctor's office; overwhelmingly they wanted the engagement to occur on a one-on-one basis.

Data users should be careful when the notification itself could reveal private or confidential information. For example, a letter mailed from an organization that supports individuals with a stigmatizing condition, such as substance abuse or HIV, could inadvertently reveal information to others, such as other members of the household.

Individual Notice Through Notice of Privacy Practices

A notice of privacy practices informs individuals about what personal information may be collected and how it may be used. Although not a notice of impending or actual use, this type of notice alerts individuals to the possibility that their data may be used in additional ways. Examples of this type of notice include the notice of privacy practices required by the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule or Terms of Use notifications on social media sites.

Individual Notice of Opt-in/Opt-out Consent

In contrast to a notice of privacy practices, notice of an opt-in or opt-out option gives individuals the notice of a consent process, as discussed in more detail below.

Community Notice

In some cases, notice is given to the community, not individuals. Different methods may be used to give notice to a community, including:

- Community meetings or town halls
- Booths at community events
- Flyers or notices posted at libraries, community centers, or government offices
- Websites or Web-based advertising
- Media stories or advertisements
- Meetings with community leaders

In cases where data about small groups of individuals are being used, more targeted notice may be needed. For example, if data use were to affect Asian women with cancer, notice could be given in a newsletter for this population, shared on blogs for or by members of this group, or posted in cancer treatment centers. Similarly, if a small geographic area is being studied, everyone on the block or in a neighborhood could be sent a letter explaining the data use that is planned.

Determining what notice should be provided

When determining the appropriate level and type of notice, data users should first determine whether laws, regulations, or agreements with a data source dictate the level and type of notice required. See "Laws and Regulations" for more information.

If no legal mandates exist, data users should consider the risk of:

- Disclosing confidential or private information
- Generating results that individuals or communities have chosen not to know or that challenge fundamental beliefs
- Stigmatizing individuals, small groups, or communities

Data users should weigh the burdens of individual notice, discussed above, against the benefits of using data. When the benefits of use are great and compelling and the cost of notice is very high or impractical, the data user may determine that individual notice is not required.

More targeted notice is warranted when individual privacy or confidentiality is at risk and when individuals can be contacted without undue expense or difficulty.

Notice can be given broadly to communities or subgroups within a community, or targeted to the individuals whose data will be used.

Engaging individuals and communities preserves the use of fetal blood spots to improve human health

When a baby is born, the hospital may collect a blood sample by pricking the child's heel. In some states, parents filed legal actions to prevent the use of these fetal blood spots for purposes that would not directly affect the child.

Researchers launched national and local efforts to understand parents' views on the issue. They learned that most parents were willing to allow the use of the blood spots for research, but parents wanted to know how the samples were being used, and they wanted the ability to limit the use.

Reflecting these preferences, states passed laws and adopted policies addressing parents' concerns about use of the blood spots. For example, in Michigan, the parents of newborns are now notified that the Michigan Biotrust hosts a website where parents can choose to limit the use of their child's blood spots through an opt-out system. If parents do not take action to opt out, the child's biological samples may be used for research.

Consent

In addition to notice, individuals may have the opportunity to choose whether their data may be used. Certain uses are required by state public health laws and do not require or offer the opportunity for individual consent. However, other situations mandate choice and consent. The HIPAA Privacy Rule and the federal regulations regarding the Protection of Human Subjects in Research, known as the Common Rule, mandate choice in many situations, as discussed in Appendix B.[2]

"Even if laws or regulations do not say how data are to be used, community stewards should assess whether ethical imperatives or the need to maintain trust require a consent process."

Consent may be required for original data collection, for example, when an individual agrees to participate in a research study. Or consent may be required for some ways of repurposing data that were not included in the original consent. For example, individuals who have consented to the use of their data to study diabetes might need to be given a chance to choose whether they want to participate in a study of correlations with mental illness or substance abuse. Even if laws or regulations do not say how data are to be used, community stewards should assess whether ethical imperatives or the need to maintain trust require a consent process. There are several approaches to obtaining consent from individuals or communities whose data are being used.

Individual Consent

Some instances of data use require individual informed consent. This requires the user to inform the individual about planned data use and to obtain the individual's consent before using the data. This type of consent is usually required in research studies, especially those where the data use has a high level of risk.

Although individual consent offers individuals the highest level of choice, it may not always be possible or feasible. For example, it may not be possible to link biological samples collected by the U.S. Army from draftees during World War II to the names of the people from whom the samples were collected and thus to obtain individual consent for use of the samples. In other cases, while it may be possible to identify the source of data, that process itself may increase the risk of violating the privacy rights or confidentiality of the person. In other cases, the cost of obtaining individual consent may be greater than the benefits.

2 Data users should take special care when requesting access to or using substance abuse treatment records, which are strictly regulated under federal law. See 42 C.F.R. Pt. 2

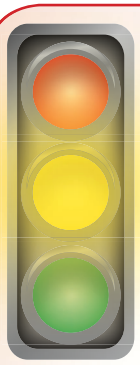
Community Consent

In cases where individual consent is not required, feasible, or warranted, data users may obtain community consent. For example, a local elected official may consent to community data being used instead of obtaining consent from individuals. This type of consent can be used when the risks to community members are relatively low, but may not be the best approach when risks to individuals or small subsets of individuals in the community are high.

Opt-in/Opt-out

In some cases, individuals may be given the choice between allowing their data to be used or not used. Opt-in and opt-out provisions usually have a default. With an opt-in approach, individuals must take action to have their data included for a particular use. With an opt-out approach, individuals' data will be available for use unless they take action to restrict or deny access to their data. Local or regional health information exchange systems typically include or exclude data based on opt-in or opt-out defaults. As noted above, these systems require notice so that individuals who are affected may exercise the choice between options.

"Local or regional health information exchange systems typically include or exclude data based on opt-in or opt-out defaults."



Cautionary Tale:

Repurposed use of blood samples

Members of the Havasupai Tribe volunteered to participate in research studies on diabetes by providing blood samples. Years later, they were surprised to find out that the researcher had used the samples to study family lineage, schizophrenia, alcoholism, and migration patterns without obtaining additional consent. In the resulting lawsuit, Arizona State University, which employed the researcher, paid the tribe a substantial financial settlement and returned the remaining samples to the tribe.

Community and Individual Engagement and Participation

Data users have an ethical, and sometimes legal, obligation to promote community and individual engagement and participation in projects that use personally identifiable, de-identified, or aggregated data and when data use could stigmatize individuals, small groups, or communities.

When data are used without appropriately engaging communities and individuals in data use decisions, trust may erode. Negative consequences of a breach of trust can have subsequent radiating effects, as shown in many case studies.

Communities can be effectively engaged at every phase of the data lifecycle and when applying stewardship principles. Engagement can be a way to protect the rights of individuals, small groups, and communities. Engagement can also help researchers or others in using data to improve health.

Mechanisms for engaging community members

Data users can engage community members in a number of ways. When determining how to engage the community, data users should think about which types of engagement would provide legitimacy for the data effort. In a politically polarized community, for example, elected officials may not be seen as representing the interests of all voters. The following briefly summarizes some approaches to community engagement.

Community Leaders

Community leaders can sometimes serve as representatives for a community as a whole. Leaders may include elected officials, leaders of community groups, leaders of religious or spiritual organizations, or even informal leaders. Use caution when using community leaders as representatives of the community, as they may not accurately represent the community's view as a whole, and they may not understand the concerns of subgroups or individuals within the community.

Focus Groups

Focus groups provide another way to engage communities, and are a good way to find out how individuals feel about an issue. Guidelines on how to run a focus group are available from: https://assessment.trinity.duke.edu/documents/How_to_Conduct_a_Focus_Group.pdf. Like engagement through community leaders, focus groups can miss issues that matter to subgroups if members of subgroups are not among the focus group members.

Community Advisory Boards

Community advisory boards are a commonly used form of community engagement. To be effective, advisory boards should represent a range of interests and subgroups within a community. One issue that must be addressed in forming community advisory boards is how members will be chosen, and whether members will be leaders of community groups, or community members who are not leaders. Some data repositories have specific requirements about characteristics of representatives who serve on advisory boards.

Community Surveys

Community surveys can be completed online, on paper, or in personal interviews. They can help data users to gather and analyze information from many people as a form of community engagement. An example of a survey to assess community members' perceptions about community health is available from: http://www.naccho.org/topics/infrastructure/mapp/framework/clearinghouse/upload/Example-Survey-CTSA-Community-Health.pdf. While a community survey can get input from more individuals, the scope of results may be limited because the scope of information is defined by the questions asked and by the characteristics of the individuals who choose to complete the survey.

The community takes the lead

In Denver, a community group called Taking Neighborhood Health to Heart is working to address a variety of health problems. The community helps to determine the questions to be asked, research to be conducted, and how and when data are released. In some cases, community members are hired to collect survey data. Because the community is an active participant in all parts of research, the initiative has learned about issues that might never have been addressed for fear that results would be used to stigmatize community members.

Stigmatization in the AIDS epidemic

In the early days of the AIDS epidemic, data suggested that Haiti was a source of the infection and that Haitian immigrants were overrepresented among the population subgroups with the disease in the United States. (See Elliott Frank, et al. "AIDS in Haitian-Americans: A Reassessment." Cancer Research 45 (Suppl 9):4619s–4620s. 1985.) The result was widespread fear of Haitian immigrants and a drop in tourism to Haiti. One of the doctors attempting to treat this population later reported that he encountered widespread mistrust because of the stigmatization. (See Ronald Bayer, Gerald M. Oppenheimer. AIDS Doctors: Voices from the Epidemic: An Oral History. New York, NY: Oxford University Press; p 28–29. 2000.)

Opportunities for engaging community members across the data lifespan

Purpose Specification

When planning projects and framing research questions, engaging the community can help data users to:

- Understand community perspectives
- Avoid mistakes that can occur when someone outside of the community makes assumptions about dynamics within a community
- Target issues that are relevant and useful to the community

Openness, Transparency, and Choice

The most important point in the engagement process occurs when implementing the stewardship principle of openness, transparency, and choice. See Openness, Transparency, and Choice for specific recommendations on community engagement.

Data Collection and Acquisition

Data users may engage communities in the data collection process, and data holders can require that those who want to use their data engage communities:

- Community members can administer surveys, which may improve participation and response rates (see The community takes the lead and the case study in Appendix C)
- Community members can provide insight into how unique characteristics of the community may affect data collection efforts (see the case study, A Refugee Community's Expectations describing the University of Maine community data project in Appendix C)
- Organizations sharing data may require that those using their data involve community advisory boards

Data analysis

Community members can explain to data users aspects of the community that may influence how data are interpreted and analyzed by individuals who do not have an understanding of community dynamics. Communities can be very helpful in reviewing findings and interpretations of findings before findings are released to the public.

Subgroup Concerns

Some data use can trigger different concerns from different communities, so data users must consider whether multiple communities or subgroups within a community should be represented. A subgroup can share a racial, ethnic, or geographic trait, or even be affected by a shared disease. Subgroup concerns can arise whether data are personally identifiable, de-identified, or aggregated.

Avoiding Stigma and Discrimination

Data users may engage communities to avoid or address concerns about data uses that have the potential to result in discrimination against or stigmatization of the community or its members. Community engagement can help data users identify areas of sensitivity or concern and be a means of addressing concerns. Data users from outside the community may not see how the data could negatively affect communities. Studies of prevalence of health issues such as sexually transmitted diseases, substance abuse, behavioral health, or genetic disorders, whether using data from medical records or public health surveillance data, may be used to identify subgroups in the population with increased risks for adverse health outcomes and have the potential to stigmatize community members. The data user should give thoughtful consideration to the use and analysis of these data to avoid stigmatizing groups or individuals.

Community engagement can also help data users to communicate findings in ways that do not stigmatize communities or subgroups, although in some cases it may not be possible to publicly release certain types of data without the risk of stigma and discrimination. Even then, community engagement in purpose specification (see below) can help data users to strike an acceptable balance between data use and the interests of research participants and communities who may want to learn from, but perhaps not publish, results.

Engaging a distinct community subgroup

The Population Study of ChINese Elderly (PINE) identified actionable concerns among older Chinese adults in Chicago, a community cohort that was less well understood. By engaging more than 20 community groups and by using multilingual staff to interview participants according to their preferred languages and dialects, the survey response rate was 91%. The result of the effort was reported in The PINE Report, which showed that members of this population are affected by medical comorbidities, physical disabilities, low health care utilization rates, psychological distress, social isolation, and elder abuse at higher rates than other older adults in the United States. The PINE Report identified opportunities for family members, community stakeholders, health professionals, and policy makers to improve the health and well-being of older Chinese adults.

Engaging the community in health information exchange

Health information exchanges enable providers to share health information across organizations and provider types to improve patient care. In some communities, concerns about privacy and confidentiality of health data have decreased information sharing through exchanges, and adversely impacted the quality of patient care. To avoid similar concerns, MyHealth Access, the Tulsa exchange described earlier, engaged the community in a 100-day planning process that involved 200–300 people. At the beginning, participants agreed to focus on the objectives of health improvement and quality. This focus allowed the community to agree on a system of privacy and confidentiality protection that permitted the flow of data needed to treat patients optimally.

Summary

- Evaluate opportunities for engaging communities and individuals at every step in the data lifecycle and across all elements of the stewardship framework
- Be aware of the concerns of subgroups within communities whose interests may be different from those of the larger community
- Consider the risk of stigmatization of communities or small groups and engage the community or individuals to determine an action plan for addressing the risk

Purpose Specification

Researchers are trained to start every inquiry by framing the question. What question is the project designed to answer? Data users should explicitly and carefully frame the question and be able to explain how the data will answer the question. This process is called purpose specification. Purpose specification helps data users reach the intended goal, regardless of the data source or type.

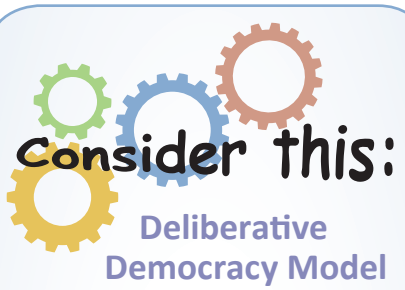
Purpose specification is relevant whether data are personally identifiable or de-identified. It is also important regardless of data source. If obtaining a data set from an entity, data users will typically need to explain the purpose for the data use. Even for data that are publicly available, explaining the purpose is important if the data use is to achieve its intended goal.

Purpose specification has many benefits:

- By requiring that data collected is carefully linked to the purpose of the project and possible follow-on projects, data collection will be targeted, focused, and thorough
- Data collection efforts that contemplate repurposing at the outset can increase efficiency while decreasing the data collection burden
- Purpose specification can help data users avoid unwelcome surprises by emphasizing the need to anticipate and plan to address negative impacts

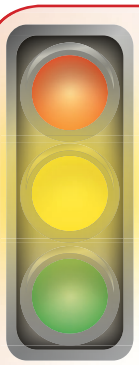
Community engagement can support the purpose specification process. Communities and individuals can help data users to understand challenges or concerns about which the data user may be unaware. Laws or regulations may dictate the purpose of data collection by government agencies, such as health surveys or infectious disease surveillance. Though overall purpose for these efforts may be broad, even these data collection efforts are usually driven by a question that the data may help to answer.

When engaged in purpose specification for a project involving original data collection, data users should anticipate and adjust for the possibility that data may be valuable for repurposing. For example, biological samples may remain at the conclusion of a study evaluating the prevalence of a vitamin deficiency. A data user, aware that samples could be used to investigate human health problems in the future, can anticipate repurposing. To address anticipated repurposing, a data user might ask for consent in the primary study for samples to be used in later studies defined in the consent.



Deliberative Democracy Model

A “biobank” collects, processes, stores, and distributes bio-specimens and related data for use in research. A biobank might include specimens of blood, saliva, plasma, or DNA. When the Mayo Clinic started biobanking and repurposing data from their electronic medical records, it adopted a deliberative democracy model that engaged community members in open dialogue for 4 days. The deliberants were provided with background materials on biobanking, biomedical research, and local efforts at Mayo. They were then given an opportunity to interact with domain experts, including scientists involved in genetics research as well as privacy advocates. The result was community support and an accepted framework for the use of biological samples and health data.



Cautionary Tale:

Repurposing data without individual or community engagement

Most newborn babies receive blood tests to determine if they have treatable medical conditions. Realizing that these blood "spots" could also be used for other purposes that would benefit public health, such as monitoring rates of genetic disorders or infectious diseases, the holders of the blood spots began to make them available for research. Parents in several states found out that biological samples taken from their babies were being used without their consent and brought legal actions. In Texas, the legal settlement resulted in the destruction of more than 5 million biological samples.

In the process of purpose specification, data users should consider the balance between defining a specific and narrow purpose or a less specific and broader purpose when using data. The advantages of a narrow scope are that the purposes are easily defined and described, so communities and individuals may be more likely to trust users and allow the desired uses of their data. However, future uses may be circumscribed. A data project that specifies a more open-ended or unknown purpose gains greater flexibility for future uses, but runs the risk that individuals may be less likely to participate because they do not understand the full extent of potential future uses for which their consent is being sought, or that future uses will surprise individuals or communities with unexpected, perhaps even unwanted, results.

Repurposed Data

Repurposed data are collected for one purpose and then used for another. Public health surveillance data collected by state health departments is repurposed when shared with communities or researchers to investigate a concern that the data may help explain. Laboratory tests performed to guide patient diagnosis and treatment are repurposed when combined with many other tests to show the prevalence of a condition in a subgroup of individuals.

When using repurposed data, users should consider concerns that may be raised by those whose data are being repurposed. The cases of the research study of the Havasupai tribe and the collection of fetal blood spots show the harm that can occur when data are repurposed without the consent of the individuals whose data are being used. The case study describing the community-based approach used by MyHealth Access shows how data users can more likely avoid problems encountered by data users who did not consider the risks of repurposing.

Public health data used by communities might have been originally collected for the purpose of controlling or preventing injury and disease, or for legal and administrative reasons, or both. For example, birth and death certificates include information useful for legal purposes (such as establishing rights to an estate), administrative purposes (establishing family benefits or ceasing benefits to decedents), or surveillance for unusual incidence of disease (such as genetic birth defects, or deaths from suicide or cancer in a geographic area). Rates of premature death, cancer, and obesity are examples of the types

of data communities can repurpose to improve community health.[3]

Users should also be aware of any limits to repurposing that may be imposed by laws governing the collection and use of the source data set or data use agreements. Laws in some states, for example, explicitly address the repurposed use of fetal blood spots. To take another example, state laws may limit the repurposing of vital statistics, such as birth and death records. However, many states have laws that allow the broad use of health care data sets to measure health care cost, quality, and access. In these states, the data steward will have a data oversight committee and data release policies that strictly govern data release and reporting uses and these are included in the DUA. In other cases, state laws or regulations allow the sharing of government health data only for specific purposes.

"Laws in some states, for example, explicitly address the repurposed use of fetal blood spots."

Tensions between data used for improving community health and for research

Purpose specification can also be used to address a tension between the goals of academic research and the goals of advancing community health. Research ethics and funding sources sometimes mandate that researchers disseminate their findings through publication or presentations at academic meetings. Communities, to the contrary, may want to use funding to improve health, while limiting dissemination of potentially stigmatizing or otherwise harmful results. Once again, community engagement in the purpose specification process can help address this tension at the outset of a project.

- At the outset of any data project, explicitly and carefully define the purpose of data collection or use of repurposed data.
- Consider how to most effectively engage the community in the purpose specification process.
- Consider and address possible adverse impacts of data use or collection.
- Be aware that data may be repurposed and design collection accordingly.
- When using repurposed data, consider how changing the

3 Community Commons offers tools to help communities use repurposed data effectively. From its website: "Community Commons is an interactive mapping, networking, and learning utility supporting broad-based and sustainable healthy communities with free access to resources for registered users." See "About" available from: www.communitycommons.org.

Using a trusted intermediary

The *Southern Illinoisan*, a newspaper, sought cancer registry data in an Illinois Freedom of Information Act request in order to see if there was a cancer cluster in an area of petroleum extraction. Dr. Latanya Sweeney, then a Professor of Computer Science at Carnegie Mellon University, and an expert in re-identification of supposedly de-identified data sets, testified that individuals could be identified using the requested data in conjunction with publicly available information because the number of cases was small. The newspaper was successful in the lawsuit and obtained the data. To avoid the suit, Illinois could have suggested disclosing the data through a trusted intermediary such as a university, which could have permitted data analysis under a promise of confidentiality in a secure setting. Communities seeking such cancer registry data might want to try this option if they encounter confidentiality concerns. *Southern Illinoisan v. Illinois Department of Public Health*, 218 Ill. 2d 390 2006).

original purpose may trigger the need for additional notice or consent or if these changes are allowed under the DUA with the data steward.

- If the project brings together academic researchers and communities using data to improve health, address any tension among academic goals, funding mandates, and community interests in protecting use limitations.

Summary

Occasionally, rare events, even in the aggregate, in conjunction with detailed local knowledge may inadvertently lead to clues or speculation about specific individuals. These effects may be in violation of explicit data use agreements or generally recognized principles of privacy.

In such cases, another strategy may be to arrange with the original data steward for some kind of trusted intermediary through which a community can analyze data in a secure data center, allowing access to the data in a controlled environment while still honoring the need to protect the confidentiality of the data in the custody of the original data steward.

Quality and Integrity

Stewardship principles require that the quality and integrity of data are managed so that they are usable for their intended purposes.

Data quality refers to the accuracy, relevance, timeliness, completeness, validity, and reliability of the data. The data collected or used for a particular purpose must have an appropriate nexus to that purpose that is timely, and as complete as reasonably necessary to answer the questions asked without bias, skewing, or other distortion. Data must be recorded or captured accurately, and it must represent what it is claimed to represent. For example, questions that are ambiguous in a survey may not yield answers that correspond to what the data user believes them to mean.

"When data sets are combined, there are risks that they may not be properly matched."

Data integrity means that the data have not been corrupted. Data users must be aware of the problem that data may be modified or otherwise garbled as they are used. When data sets are combined, there are risks that they may not be properly matched. Therefore, the combined data may no longer accurately reflect the sources.

It is seldom possible or necessary to have perfect data, but stewards should consider and make a judgment about whether data accurately and adequately measure what is being studied, and if the data can be trusted. The data stewards of large public health databases should provide detailed documentation about the underlying data and its limitations, and should be consulted to validate and review findings prior to public release of reports or statistics derived from these data sets.

Data Quality and Integrity Through the Lifecycle

Review of the Literature

Data users should research and evaluate what has already been done; doing so helps to ensure the quality of data and can answer the following questions:

- Is further data collection needed, or is the necessary information already available?
- If others have addressed the issue in a different population, can a proven methodology be used rather than starting from scratch?
- What methodologies have failed to work?

By starting with a scientific literature review, the data user can avoid duplicating effort and avoid others' past mistakes.

Data Collection, Data Entry, and Data Cleaning Processes

To ensure data quality, users should assess that original data are collected in accordance with generally accepted procedures, and that sources of repurposed data are trustworthy. For example, a trustworthy data source would be able to provide assurances about how data were collected, entered into a database, and stored. The Data Quality and Integrity Checklist (Appendix D) outlines the steps for users to follow.

"...a trustworthy data source would be able to provide assurances about how data were collected, entered into a database, and stored."

Analysis

Data analysis should be conducted by trained and experienced individuals or entities. If an organization lacks internal experience, it may consider associating with researchers who are interested in the issue being studied.

Reporting Results

Results, whether published in a journal or report or used within an organization for internal purposes, should accurately describe the results findings of the analysis and should avoid bias.

Special Consideration for Merged Data Sets

Data users sometimes merge data from two or more sources to gain enriched data that is more useful than either data set alone. However, data users must be careful to combine data sets where the measures use the same populations, standards, and scale, so that they are not comparing apples and oranges but using data to make valid inferences.

Examples of Merged Data Sets

- The results of a survey of nutritional habits of adolescents, administered by different school districts in different cities in a state, could be combined to increase the study's statistical power.
- Two different data sets could be combined to better understand a phenomenon. For example, obesity rates obtained from government sources could be combined with a map of safe walking routes to consider whether lack of safe walking routes is associated with higher rates of obesity.

Validity of Merged Data Sets

When two or more data sets are combined, users should ensure that a merger or aggregation is valid, and that the data retain integrity. In determining validity, data users should ask:

- Are the populations the same for the different data collection efforts?
- Do survey questions and response categories match?
- Might differences in survey administration dates affect survey results?
- What were the survey sample designs?

Many of the issues involved in determining if survey data can be combined and how they should be combined are substantive, and require consideration by subject-matter experts. These issues should be resolved before any statistical consultation takes place.[4]

"Might differences in survey administration dates affect survey results?"

Summary

- Ensure that data quality and integrity are maintained throughout the data lifecycle, as outlined on the Data Quality and Integrity Checklist in Appendix D.
- Before merging data sets, consider how the merger will affect data quality and integrity.

4 For a detailed discussion on how to evaluate the validity and integrity of merging data sets, see the U.S. Department of Health and Human Services, Office of the Assistant Secretary for Planning and Evaluation's Data on Health and Well-being of American Indians, Alaska Natives, and Other Native Americans: Data Catalog, Contract No. 233-02-0087, Appendix B: Data Set Aggregation, B-1 (Dec. 2006), available from: http://aspe.hhs.gov/hsp/06/catalog-ai-an-na/report.pdf.

Security

"Security is particularly important for personally identifiable data that are private or confidential."

Securing data means protecting the data's confidentiality, integrity, and availability. Good security protects data from loss of control, and, therefore, potential unauthorized access, damage, or manipulation. Security safeguards may be technical, administrative, or physical controls and can range from using locks on an office door and procedures for handling paper forms to the use of sophisticated encryption software. Security is particularly important for personally identifiable data that are private or confidential.

Some of the primary threats to loss of data include using weak passwords; failure to back up electronic data; infection by viruses or malware; and loss of portable electronic devices, such as smart phones, thumb drives, and laptop computers. Employees can increase the likelihood of security incidents by either failing to follow policies and procedures designed to protect data security, or by deliberately taking, altering, or destroying data. Even paper can be at risk—for example, completed surveys could accidentally be placed in a recycle bin during an office cleanup.

Responsible data security includes these steps:

- Evaluate anticipated risks
- Develop a plan to reduce anticipated risks
- Re-evaluate risks periodically

Elements of a security plan could include:

- Identification of major risks
- Adoption of methods to secure paper documents
- Password protection for access to computers, networks, and electronic devices
- Encryption of data stored on removable devices such as laptops, tablets, or phones, ensuring data cannot be accessed if the computer is lost or stolen
- Automated backup processes to protect against accidental data loss
- Training for employees on security measures
- Signed confidentiality agreements from all staff collecting, managing, analyzing data

As with the question of notice, data users must assess the need to secure data, and the costs of doing so, against the risk of data loss, inappropriate access, or manipulation.

Ways to improve data security

- Physical
  - Install locks on cabinets or rooms where paper records are stored
  - Keep records away from areas vulnerable to damage in a flood
  - Protect electronic storage facilities against break-ins or destruction
  - Back up data with off-site storage capabilities
- Administrative
  - Run a risk analysis
  - Set up policies and procedures for accessing paper records, disposing of data, or adding new equipment on a network
  - Train those with access to sensitive information in data security
  - Require robust passwords
  - Control who has access to view or change the data
  - Conduct due diligence on employees who handle data
  - Implement an incident response program
- Technical
  - Maintain logs of system access and unauthorized extraction of data
  - Add encryption
    - Specific elements in a data set
    - Data set as a whole
    - Devices that allow access to the data set, such as laptop computers
  - Implement monitoring to scan for and identify cyber attacks

For more detailed information about security, see the National Institute of Standards and Technology guides on assessing and maintaining data security,[5] which are useful for nonfederal organizations. The Office for Civil Rights of the U.S. Department of Health and Human Services also publishes security guidance in plain language[6] for entities covered by the HIPAA Security Rule, which nonetheless is instructive for organizations not covered by that rule.

Role of De-identification in Data Security

De-identification is a process where personal identifiers such as name, address, telephone number, or date of birth reduce the risk that private or confidential information will be disclosed. The process of de-identification and protection from re-identification are addressed in the next section.

"De-identification is a process where personal identifiers such as name, address, telephone number, or date of birth reduce the risk that private or confidential information will be disclosed. "

5 A list of guides published by the National Institute of Science and Technology's Computer Security Resource Center is available from: http://csrc.nist.gov/publications/PubsSPs.html.

6 Educational materials from the Office for Civil Rights about the HIPAA Security Rule and other sources of standards for safeguarding electronic protected health information include the HIPAA Security Information Series, available from: http://www.hhs.gov/ocr/privacy/hipaa/administrative/securityrule/securityruleguidance.html. In particular, HIPAA Security Series 1: Security 101 for Covered Entities gives an overview of basic concepts, and HIPAA Security Series 7: Security Standards: Implementation for the Small Provider describes basic topics for data users. Other available resources include "Privacy and Security Training Games"(http://www.healthit.gov/providers-professionals/privacy-security-training-games); "Guide to Privacy and Security of Electronic Health Information" (http://www.healthit.gov/sites/default/files/pdf/privacy/privacy-and-security-guide.pdf); "Security Risk Assessment Tool" (http://www.healthit.gov/providers-professionals/security-risk-assessment); and "Your Mobile Device and Health Information Privacy and Security" (http://www.healthit.gov/providers-professionals/your-mobile-device-and-health-information-privacy-and-security).

De-identification

De-identification is the process of removing or obscuring any directly or indirectly identifying information from data in a way that minimizes the risk of unintended disclosure of individuals' identity and information. By removing directly identifying elements and otherwise treating data through de-identification, released information can be both confidential and useful for legitimate purposes.

Good de-identification practices reduce risks of re-identification to a level judged acceptable given the data's sensitivity. Using de-identified data whenever possible is a strong privacy practice, because it reduces risks of a data breach and other violations of personal privacy. Data de-identification makes it very hard to link data to a specific individual, allowing the study of a variety of sensitive issues while greatly reducing the risk of disclosing personal or confidential information. Aside from organizations that must follow HIPAA de-identification methods, no standard, universally adopted de-identification method is used throughout health care.

"Good de-identification practices reduce risks of re-identification to a level judged acceptable given the data's sensitivity."

Identity and Attribute Disclosures

There are two areas of concern regarding re-identification: The first is identity disclosure, which happens when an outside party can assign an identity to a record in a disclosed data set; the second concern is attribute disclosure.

Attribute disclosure allows an outside party to attribute characteristics to someone in a data set even if he or she has not been individually identified. This form of disclosure is of primary concern in summary data releases, and it may arise from the presence of empty cells either in released tables or linkable sets of tables. The presence of a zero cell within a table could allow an outside person to infer that no one in the particular category had the characteristic in question. This could be very sensitive information. For example, the zero cell could indicate lack of control of blood glucose levels, and, by inference, that no one in a specific category of diabetes patients defined by race and sex had good control of their blood glucose levels.

If the opposite is true—for instance, a cell has 100% of a particular subgroup in a sample showing a specific attribute—then membership in the subgroup implies having that attribute. For example, if all of the homosexual men in a sample are positive for Hepatitis C, then any homosexual man in the sample can be assumed to have Hepatitis C.

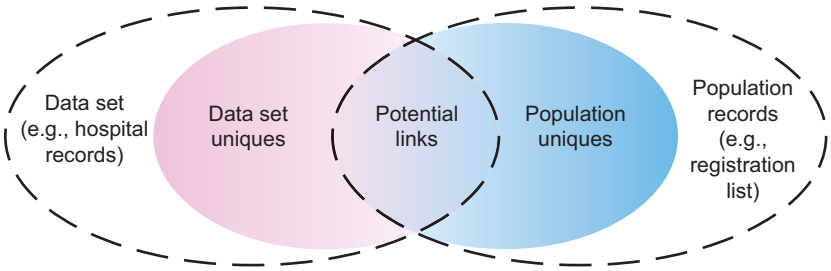
Simple De-identification

De-identification in its simplest form means deleting a patient's name from the associated health record. However, even before the advent of computer databases, this simple form of de-identification would have been insufficient to maintain confidentiality. Learning how re-identification attacks happen provides some answers:

Even when an administrator removes all of the data fields he or she thinks might be uniquely identifiable from a data set, an attacker may still be able to unlock the identity of the subject of a record by discovering pockets of uniqueness remaining in the data. This type of re-identification is possible because, even without a specific identifier, certain combinations of values may be so rare that they create a "fingerprint" pointing to only one person. A re-identification attack attempts to locate the unique fingerprints in a de-identified data set, and then search for that same fingerprint in another data set containing unique identifiers. This technique is best shown using a Venn diagram:

"This type of re-identification is possible because, even without a specific identifier, certain combinations of values may be so rare that they create a 'fingerprint' pointing to only one person."

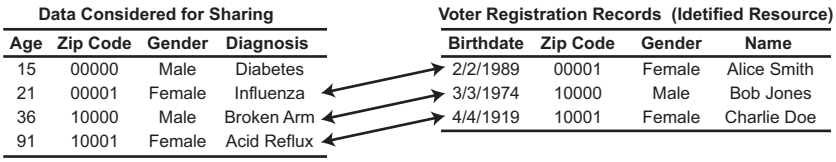
Looking for Unique "Fingerprints" in a Database[7]



This process of re-identification can be as simple as doing a reverse phone number lookup on a data set containing phone numbers. In a more complex form, the re-identification attack might identify a health record with a combination of age, zip code, and sex that is unique in the data set, and then cross-reference that information with a voter registry to determine the one individual in that zip code of that sex who was born on that day. De-identification tries to protect against this external linkage via uniqueness.

7 See "Understanding HIPAA Privacy," published by the Office for Civil Rights, U.S. Department of Health and Human Services, in Health Informat ion Privacy: Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule, available from: http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html.

Re-identification Using Public Records[8]



Data Considered for Sharing

Age	Zip Code	Gender	Diagnosis
15	00000	Male	Diabetes
21	00001	Female	Influenza
36	10000	Male	Broken Arm
91	10001	Female	Acid Reflux

Voter Registration Records (Idetified Resource)

Birthdate	Zip Code	Gender	Name
2/2/1989	00001	Female	Alice Smith
3/3/1974	10000	Male	Bob Jones
4/4/1919	10001	Female	Charlie Doe

De-identification methods must not only attempt to remove any information that would be personally identifiable, but also manipulate the data set to ensure that it contains no unique fingerprints.

Individual-level De-identification

Data users can de-identify individual records through a number of methods. The most common are suppression, generalization, and distortion.

"Use of de-identified data whenever possible is a good privacy practice as it reduces risks of a data breach and other violations of personal privacy."

Example: Data Set

Age (years)	Gender	ZIP code	Diagnosis
15	M	00000	Diabetes
21	F	00001	Influenza
36	M	10000	Broken arm
91	F	10001	Acid reflux

Suppression occurs when information is completely removed from the data set. Direct identifiers such as names and Social Security numbers are common examples of an individual's data that are completely suppressed. Some data such as birth dates and zip codes, however, cannot be completely suppressed without destroying the utility of the data set.

Example: Data Set—Suppressed

Age (years)	Gender	ZIP code	Diagnosis
	M	00000	Diabetes
21	F	00001	Influenza
36	M		Broken arm
	F		Acid reflux

Where complete suppression is impractical, data are often generalized. In generalization, a particular variable, such as age, is divided into broader categories, such as 5-year age spans. Generalization is often extremely effective at balancing utility and privacy in a data disclosure.

8 Id.

"Data users can de-identify individual records through a number of techniques."

Example: Data Set—Generalized

Age (years)	Gender	ZIP code	Diagnosis
< 21	M	00000	Diabetes
21 ≤ 34	F	00001	Influenza
35 ≤ 44	M	10000	Broken arm
> 45	F	10001	Acid reflux

Distortion may also be used to de-identify data, but with health data, distortion often destroys the reliability of the data for use in drawing effective findings.

De-Identification Through Aggregation

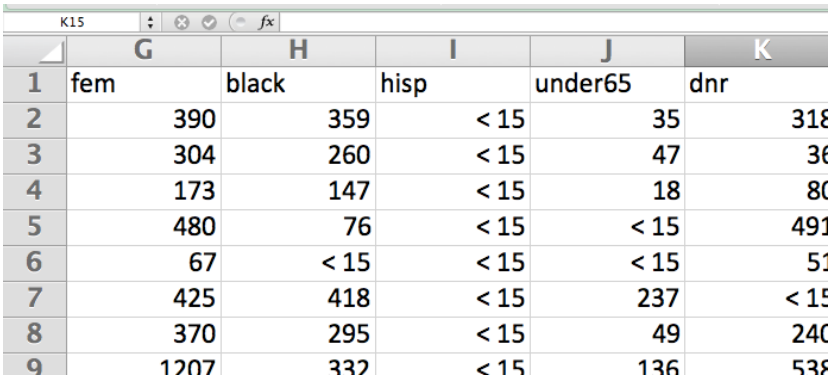
Aggregation is another way to de-identify data. Instead of removing identifiers from individual-level data, data can be combined into aggregate, or statistical, reports. This form of de-identification can be particularly effective at maintaining utility while protecting the data's confidentiality. However, the risk of inadvertent attribute disclosure remains. For example, the following table logically implies that all Hispanic females enrolled in the Healthyville School District during the 2014–2015 school year and included in the survey used illicit drugs.

Example: Data Set—Inadvertent Attribute Disclosure

2014–2015 Healthyville School District Drug Usage Survey			
	No drugs	Illicit	Illegal
White male	85	40	15
White female	90	12	7
Black male	45	15	8
Black female	50	11	13
Hispanic male	10	5	7
Hispanic female	0	3	0

When releasing aggregate or statistical reports, one effective strategy is to avoid small "cell" counts. When a cell in aggregated data is small, it increases the risk of re-identification. For example, when a data set contains health data representing thousands of patients, but only four patients are affected by a specific type of cancer, those four patients are at high risk of being identified. In the aggregated data shown in the following figure, the number of persons of Hispanic origin is so small that reporting the number of those individuals raises the risk of re-identification.

Aggregated Data



	G	H	I	J	K
1	fem	black	hisp	under65	dnr
2	390	359	< 15	35	318
3	304	260	< 15	47	36
4	173	147	< 15	18	80
5	480	76	< 15	< 15	491
6	67	< 15	< 15	< 15	51
7	425	418	< 15	237	< 15
8	370	295	< 15	49	240
9	1207	332	< 15	136	538

The risk of re-identification also increases when data are combined from more than one source, or when data represent members of a small group of people, whether they are members of an ethnic or racial minority, or members of a group suffering from a specific illness.

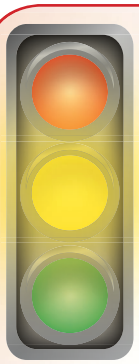
Even when aggregation is used, and even if small cells are not reported, some risk of re-identification may remain. If this is the case, data users should seek expert advice for assistance in methods to further reduce these risks.

In addition, data users can use data use agreements, discussed in the following "*De-identification, Limited Data Sets, and Data Use Agreements* " section, to limit attempts at re-identification. Another approach is to ask individuals whose data are to be used if they would consent to data use even if there were a risk of re-identification.

Although the risk of re-identification may not be eliminated, the risk may be outweighed by the benefits of using health data. Data users should explicitly address the tension between the desire to maintain confidentiality and privacy and the desire to use data to advance public health. The data steward must consider a series of tradeoffs, including the application of rigorous statistical and data management controls to reduce the risk of re identification, while preserving as much data utility as feasible.

State health data organizations that maintain hospital discharge databases apply a layered approach to protecting the data sets they release, following standards adopted by the National Association of Health Data Organizations (NAHDO). This methodology reduces the probability of unique re-identification of individuals through statistical and technical modifications that alter the data. De-identification combined with data management measures (such as data oversight boards, training and education of users, or penalties for misuse), and information technology solutions (such as

"This form of de-identification [aggregation] can be particularly effective at maintaining utility while protecting the confidentiality of the data."



Cautionary Tale:

Small cell sizes

An academic used state vital records data from death certificates to study cause of death from a variety of causes. This researcher was able to identify one individual because of small cell size. As a result, the government agency that supplied the data decided to increase the suppression criteria from 5 to 10. They put in place a system where an automatic check is performed in the background before results are reported back to a researcher to check for cell sizes smaller than 10. Now, if someone runs an analysis for which any of the cell sizes are less than 10, the cell will come up blank or just indicate "<10."

encryption), are methods that may help to manage the risk of release while making relevant health care information more available to data users.

Quantifying and Evaluating the Risk of Re-identification

Evaluating risk of re-identification can be a very technical process that requires substantial expertise, but community data users can use certain general principles as a guide. The most important factor to consider is the number of individuals who share a certain set of characteristics. Name, address, and telephone number are obvious examples of data elements that can reveal the identity of a person, but other data elements may be less obvious.

Communities should be aware that merging data sets, in particular, may increase the risk that individuals or small groups could be identified. Merged data sets raise concerns when people would not expect the data to be combined (for example, correlations among prescriptions filled, food purchases, and method of payment for food that could be obtained from private supermarket data); when analysis of the combined data sets may have negative consequences for those whose data are used; or when merger raises the risk that private or confidential data may be disclosed.

Good data stewardship practices require evaluating the re-identification risks for new mergers of de-identified data sets, and for all new uses of de-identified data sets. The Office for Civil Rights of the U.S. Department of Health and Human Services provides guidance[9] on how HIPAA-covered entities can evaluate risk of re-identification, but community-based data users should not undertake this process without expert guidance.

De-identification, Limited Data Sets, and Data Use Agreements

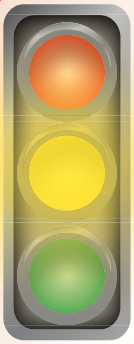
The HIPAA Privacy Rule requires data use agreements (DUAs) when researchers use "limited data sets." A limited data set is created from protected health information by removing all identifiers except certain information about dates and locations. Users obtaining de-identified data sets also may need to enter into a DUA with the entity supplying the data, to

9 See Health Information Privacy: Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule, published by the Office for Civil Rights, U.S. Department of Health and Human Services, available from: http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html.

promise to protect the data against re-identification or to make additional privacy and security arrangements. Data stewards of state data sets may be subject to laws prohibiting attempted re-identification of records, and many DUAs impose penalties for noncompliance with DUA requirements. Communities that engage in the collection of original data may share de-identified data with other organizations, and when doing so, should use a DUA to make clear the expected arrangements for use of the data, including limiting attempts to re-identify de-identified data.

Summary

- De-identification can be used to limit the risk that individuals' confidential or private data will be disclosed.
- Two types of de-identification are:
  - Individual de-identification
  - Aggregation
- Data users can use a number of strategies for limiting the risk of re-identification, such as:
  - Suppressing small cell counts
  - Grouping variables that could make re-identification easier
- Data use agreements that prohibit attempts to re-identify individuals can add a layer of protection to other strategies for protecting confidentiality and privacy.
- When de-identification interferes with the purpose of the data use, individuals can be asked if they accept the risk of re-identification.



Cautionary Tale:

Washington State Hospital Discharge Data

A researcher purchased hospital discharge data from the state of Washington. Although the data set did not include patient names, the researcher was able to corroborate highly sensitive information about specific individuals by linking publicly available information from newspaper reports about accidents to the information contained in the data set.

Washington did not use the NAHDO Guidelines for release, which recommend a layered approach of statistical, management, and regulatory controls. Washington learned from the experience and put in place a system using data use agreements that, among other things, requires researchers accessing data to agree not to try to re-identify individuals in the de-identified data set.

Appendix A: Definitions

The following definitions explain how terms are being used in the Toolkit, although the definitions are similar to other common uses.

Confidentiality

The treatment of information that a person has disclosed in a relationship of trust with the expectation that it will not be passed on to others in ways that are inconsistent with the understanding of the original disclosure without permission.

Consent

A process through which a community or individual gives permission for data to be collected or used by a specific entity for a specific purpose.

De-identified Health Data

Health data about an individual that has had identifiers, such as name, address, telephone number(s), and date of birth removed. For HIPAA-covered entities using protected health information (PHI), the HIPAA Privacy Rule governs the specific data elements that must be removed to create a de-identified data set.

Health Data

Information about the health of specific individuals, such as blood pressure, or about subgroups of individuals, such as children under 5 years old with asthma living in a specific zip code, or about a community, such as the number of residents with stage 4 adenocarcinoma of the colon.

HIPAA

Health Insurance Portability and Accountability Act. The part of HIPAA that most people have encountered is the Privacy Rule, which gives certain rights to individuals—for example, to obtain copies of their medical records—and imposes duties on health care providers, their business associates, and insurance companies or other payers to maintain the privacy and confidentiality of patient information.

IRB

Institutional Review Board. A structure created by the "Common Rule" for the Protection of Human Subjects in Research, that ensures that research involving people meets legal and ethical requirements. Federal and state laws and regulations determine what research must be approved by an IRB.

Notice

Information given to the community or individuals about how their data may be used.

Protected Health Information or PHI

Refers to information about an individual that is subject to the HIPAA Privacy Rule. PHI receives specific legal protections under the HIPAA Privacy Rule.

Stewardship

Health data stewardship is a responsibility, guided by principles and practices, to ensure the knowledgeable and appropriate use of data derived from individuals' personal health information.

User of Community Health Data

Entity within a community that collects, manipulates, stores, analyzes, or disseminates data to improve the health of a community, or of subgroups or individual members of the community.

Appendix B: Federal and State Laws

Many federal and state laws and regulations could affect community level data use, but two sets of federal regulations are most likely to affect local efforts to use data. Because there are 50 states with 50 sets of laws that may affect data use, the Toolkit does not address state law, but data users should learn about the laws in their jurisdiction.

The Department of Health and Human Services (HHS) regulations on the Protection of Human Subjects are found in the U.S. Code of Federal Regulations, Title 45, Part 46 (45 CFR 46). These regulations govern human subjects research across a range of settings, including research done by universities, state and local governments, and nonprofit organizations. Research activities covered under 45 CFR 46 must be approved by an Institutional Review Board (IRB). This Toolkit provides guidance to data users to help them determine if data use requires IRB oversight.

The Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule may also apply to entities disclosing data when communities are seeking access to health data, and it may be useful to understand the duties and limitations of the entities from which communities want to obtain data.

This Toolkit does not give data users everything they need to know about HIPAA or human subject protection rules and regulations. Instead, the goal is to alert data users to situations when they need to ask for further guidance from attorneys, or compliance experts to ensure that data use complies with major federal regulations that govern health data use and data collection efforts.

General Principles

Although the HIPAA Privacy Rule and rules governing human subjects research may not apply to community-level use of data to improve health, the underlying principles of these laws and regulations can be instructive to data users. These laws and regulations were developed to respond to concerns about perceived and actual harm resulting from data use in the past. If a data user finds itself in an ungoverned area, he or she should think about the types of protections and inquiry required of data protection and sharing under the HIPAA Privacy Rule and human subject protection laws and regulations. These protections may sometimes put limits on data sharing that would be unduly burdensome when using data to promote community health; they may also be less restrictive than some communities would want when the risk of harm to small groups or individuals is very high.

How the Regulatory Structure of Data Can Allow Community User Access to Data

By learning how data are regulated, communities may be more effective in accessing data needed to promote community health. For example, a community that understands which data are regulated by HIPAA may be more confident in reaching out to health information exchanges or providers to request data. Similarly, communities may be more willing to engage with researchers from a local college or university if they understand the role of Institutional Review Boards for the Protection of Human Subjects in Research. The final section of this tool kit gives community data users an introduction to these systems.

Human Subjects Research

A brief summary of the regulation at 45 CFR 46, Protection of Human Subjects, also known as the Common Rule, is given to prompt community groups using health data to think about whether projects must comply with this federal regulation. The most authoritative primary source of information about federal human subjects regulations is found on the website for the Office of Human Research Protections of the U.S. Department of Health and Human Services: http://www.hhs.gov/ohrp/index.html.

Other federal and state laws and regulations may impose requirements on data collection and use. For example, efforts to test interventions or collect data in the schools may be affected by education laws and regulations.

Federal law defines **research** as "*a systematic investigation*, including research development, testing, and evaluation, designed to develop or contribute to generalizable knowledge."

A *human subject* is "*a living individual* about whom an investigator doing research obtains either

- data through *intervention* or *interaction* with the individual , or
- identifiable *private information*."

Interventions include physical procedures, such as collecting a blood sample, or manipulating the person's environment—for example, changing the placement of fruits and vegetables in a local market as part of a project to measure whether the change affects the amount of fruits and vegetables purchased.

Interactions include any communication or contact between a data collector and the person, which occurs, for example, when a data collector interviews a person.

Private information is information about people collected in a place where the person would expect privacy, such as inside their home. An observation of mothers with their children in a public playground would not be private information. But private information does include information given by a person for specific purposes that is expected to remain private (for example, a medical record). Information that a person gives to a reporter would not be private information. If the information is not linked to a specific person who is or may be identified, it is not considered private information under 45 CFR 46.

Systematic investigation

A systematic investigation is a plan to collect and analyze data for answering a question. Systematic investigations include:

- Medical chart reviews
- Surveys and questionnaires
- Interviews and focus groups
- Analysis of biological specimens
- Epidemiological studies
- Psychological or sociological experiments
- Analysis of repurposed data

Generalizable knowledge

Data collection that is "designed to develop or contribute to generalizable knowledge" includes efforts to set up a knowledge base that can be applied to other communities. For example, a community group may want to influence policy about school nutrition. They design a project where their members interview students across a random sample of schools across the city about their food choices in school cafeterias. They expect that the results can be presented to the news media, that they might be used to change laws on student nutrition, and that they might be presented at a national conference. This project would likely be considered research.

Some activities are usually not considered research:

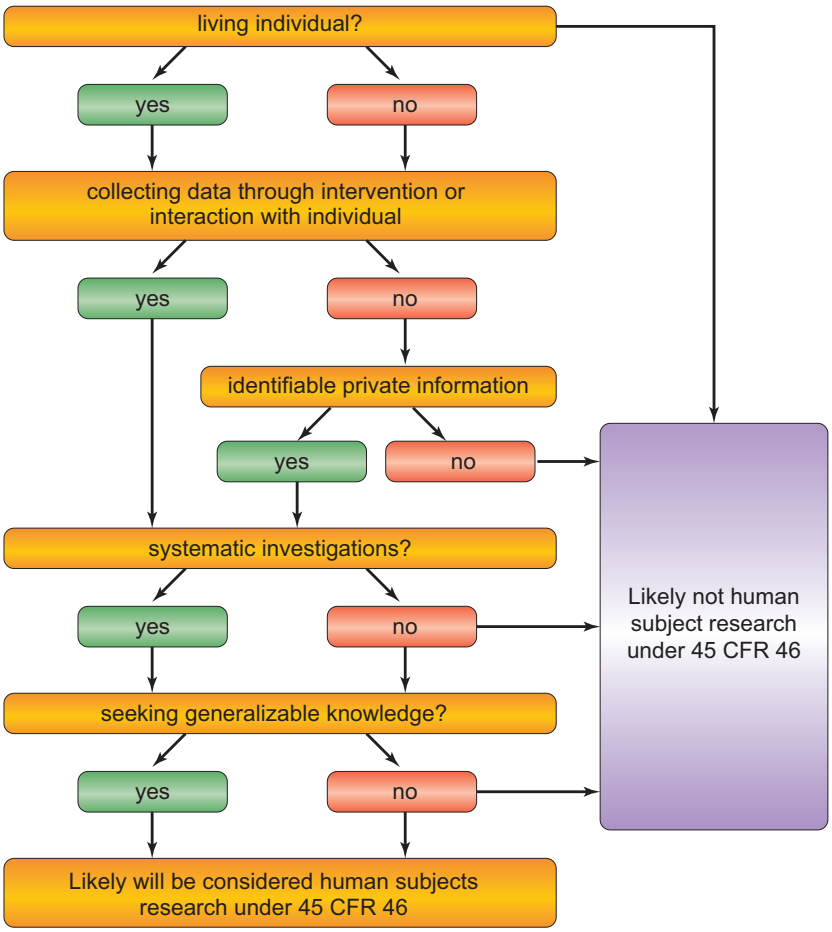
- Biographies or oral histories documenting past events
- Employee or student evaluations
- Data or evaluation collected for use internal to an organization that will not be shared with the public
- Quality improvement activities that will not be shared with the public

An IRB may need to review a proposed project to ensure that these activities are not research under 45 CFR 46.

Next Steps

Community data users who determine that a project is or may be research with human subjects should consult an IRB or compliance officer to determine what they must do to comply with any laws and regulations governing their project.

Is a project "human subjects research"?

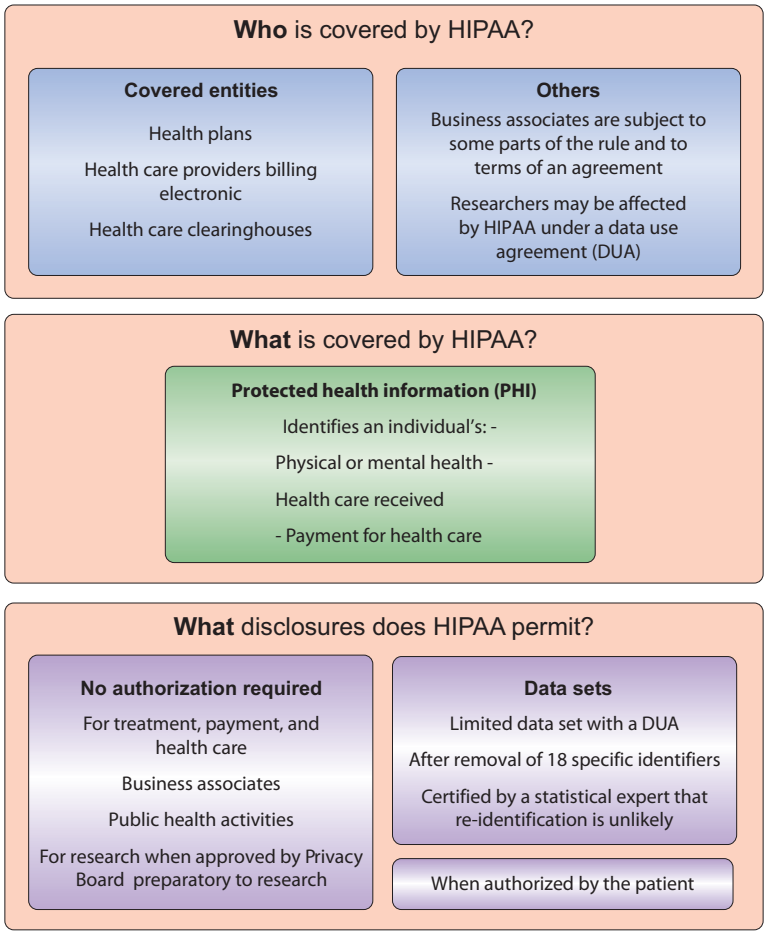


Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule

A brief summary of the HIPAA Privacy Rule is provided to prompt community groups using health data to consider whether they may be covered and to understand the duties of entities providing data to them[1].

Health data users should know:

- Individuals and organizations covered by the HIPAA Privacy Rule
- Information protected by the HIPAA Privacy Rule
- Disclosures of information allowed by the HIPAA Privacy Rule
- Notification that must be given to individuals whose data are being shared



1 A comprehensive, authoritative summary of the HIPAA Privacy Rule may be obtained from the Office for Civil Rights, U.S. Department of Health and Human Services, Summary of the HIPAA Privacy Rule, at http://www.hhs.gov/ocr/privacy/hipaa/understanding/summary/. The full text of the HIPAA Privacy Rule can be found at 45 CFR **Part 160** and Subparts A and E of **Part 164**.

An entity asking to use data from a HIPAA-covered entity (broadly speaking, health care providers, insurers, and health care clearinghouses) may need more information than is included in this Toolkit.

HIPAA Privacy Rule and Research

The Privacy Rule specifies when a covered entity may share an individual's data without an authorization for release from the patient. The following is provided to help data users understand the limitations on data sharing by covered entities. The covered entity is allowed to share patient data only when doing so complies with the HIPAA Privacy Rule. The Privacy Rule addresses access to protected health information, not human subjects research; projects using protected health information from covered entities are governed by the Privacy Rule and regulations protecting human subjects in research.

De-identified Data

Researchers may be able to access de-identified patient data from a covered entity. The Privacy Rule does not restrict the use or disclosure of ***de-identified data, but there is no requirement that a covered entity disclose de-identified data.*** Data are considered de-identified if the 18 identifiers listed below are excluded from the data used for research and the covered entity does not know that remaining information can be used to identify the individual, or if a qualified statistician determines that the data are de-identified.

Privacy Rule De-identified Data Elements

To create a de-identified data set from HIPAA-protected health information, a covered entity must remove the following identifiers:

- Names
- *Geographic subdivisions smaller than a state
- *Dates
- Telephone numbers
- Fax numbers
- E-mail addresses
- Social Ssecurity numbers
- Medical record numbers
- Health plan beneficiary numbers
- Account numbers
- Certificate/license numbers
- Vehicle identifiers and serial numbers, including license plate numbers
- Device identifiers and serial numbers
- Web universal resource locators (URLs)
- Internet protocol (IP) address numbers
- Biometric identifiers, including fingerprints and voiceprints
- Full-face photographic images and any comparable images
- Any other unique identifying number, characteristic, or code, unless otherwise permitted by the Privacy Rule for re-identification

*Identifiers marked with an asterisk may be included in a limited data set.

Limited Data Set

Recognizing that de-identified data may be needed for research to improve health, the HIPAA Privacy Rule allows covered entities to use or share a limited data set. A limited data set excludes most, but not all, elements excluded in a de-identified data set. Specifically, certain dates and geographic data may be provided in a limited data set. A covered entity may use or disclose a limited data set only for research, public health, or health care operations. In addition, the covered entity must have a data use agreement when sharing a limited data set.

Relationship Between HIPAA Privacy Rule and Protection of Human Subjects in Research

Meeting the Privacy Rule's requirements for receiving health data from a covered entity does not relieve an organization of meeting requirements imposed on research involving human subjects. An organization planning to use data from a covered entity should consult with an Institutional Review Board or compliance officer to determine additional requirements that other federal or state laws or regulations may impose.

Appendix C: Case Studies

eMERGE Network

The eMERGE network is studying the relationship between genome-wide genetic variation and common human traits. The eMERGE network has emphasized privacy and ethical data use.

Members of the eMERGE network have used a variety of methods for engaging communities in discussions about the use of individuals' genetic samples. In Phase 1, four of five sites used Community Advisory Boards; three of five sites used focus groups; and fewer than three used telephone surveys, consensus panels, deliberative engagement, Web surveys of different populations, interviews, or newsletters.

Just as different network members used different ways to engage the community, they have different approaches to protecting individual privacy and confidentiality. The eMERGE network is continuing to work to define what it means to de-identify biospecimens, biological data, and clinical information.

Vanderbilt

Vanderbilt's system involved a Web survey of 4,037 individuals and a Community Advisory Board, set up to ensure that the community had a voice. Board members worked with members of the eMERGE network at Vanderbilt and brought information back to the community. It initially consisted of 12 individuals who represented interests including parenting, church groups, civic communities, and education. Board members were not expected to have educational or genetics backgrounds. Vanderbilt found community board members to be inquisitive and active participants. They were not passive; instead, they wanted to know about what the eMERGE network was doing, and they wanted to give recommendations.

Vanderbilt also found that community boards alone were not enough: People in the community needed a specific person to talk with about the project. That focal person, sometimes called an ombudsman, can explain the organization's accountability policies and procedures when working with the community and ensure that concerns reach the right person.

Members of the eMERGE Network have found that community engagement has been "a lifesaver."

Although Vanderbilt's Institutional Review Board did not view the project to be "human subjects research" (see Legal), they added more layers of oversight, including evaluation by the university's Ethics Committee and three oversight

boards: Ethics, Scientific, and Community Advisory Boards. Their de-identified repository allows individuals to opt out of participation. In addition, researchers using eMERGE data must register each study separately, and alert researchers when their data use may violate policies and the intent of the persons in the community whose data are being used.

Sources:

Bradley Malin, Ph.D., Vanderbilt University (testimony and correspondence)

eMERGE website: http://emerge.mc.vanderbilt.edu/

Mayo Clinic

When the Mayo Clinic started biobanking and reuse of electronic medical records, it adopted a deliberative democracy model. The model engaged people in the community in open dialogue for four days. The deliberants were given background materials on biobanking, biomedical research, and local efforts at Mayo. They were then given an opportunity to interact with domain experts, including scientists involved in genetics research, as well as privacy advocates.

Participants debated the issues and formulated specific recommendations about how Mayo should address notice, consent, and privacy within its biobanking and medical record reuse system.

Sources:

eMERGE website: http://emerge.mc.vanderbilt.edu/

McGuire AL, Basford M, Dressler LG, Fullerton SM, Koenig BA, Li R, McCarty CA, Ramos E, Smith ME, Somkin CP, Waudby C, Wolf WA, Clayton EW. Ethical and practical challenges of sharing data from genome-wide association studies: The eMERGE Consortium experience. Genome Res 2011 21(7):1001-7.

Newborn Blood Spots

Almost every baby born in the United States is screened for a range of diseases via the taking of a small amount of blood shortly after birth. Parents have been routinely told that the blood spots are used for diagnosis and quality improvement. Over time, however, researchers realized that the blood spots could be used for biomedical research that could potentially benefit public and individual health. Officials in some states allowed blood spots to be used for research purposes without first notifying parents about the repurposing of the blood spots.

When some parents learned that the samples were stored long after the blood spots were used to diagnose diseases in newborns, and were later used for research without consent or notification, they brought lawsuits against states, academic institutions, and researchers. Although a case in Minnesota was dismissed, a Texas case was settled after the parties reached an agreement to destroy 5.3 million newborn blood spots. The destroyed samples were potentially a valuable source of information about genetic variation, infectious disease, and other public health challenges.

The U.S. Department of Health and Human Services engaged researchers to evaluate parents' preferences about future use of newborn blood spots. The researchers reported that most parents approved of using the samples for research, but they wanted to be notified of the possible use. Some asked for the ability to opt out of research.

For samples collected after April 30, 2010, parents of children born in Michigan can opt out of research on behalf of their children, but if they do not opt out, the biological samples default to an "opt in" status. Michigan BioTrust has created a website where parents can learn more and complete the process of opting in or out of research. This website is a good example of how data users can promote openness, transparency, and choice.

Sources:

Botkin JR, Goldenberg AJ, Rothwel, E, Anderson RA, Lewis MH. Retention and research use of residual newborn screening bloodspots. Pediatrics. 2013; 131(1):120–7. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3529945/.

Michigan Department of Community Health. Biotrust consent options. Available from: http://www.michigan.gov/mdch/0,1607,7-132-2942_4911_4916_53246-244016--,00.html.

Community Engagement on the Community's Terms

In tribal communities, leaders may be older individuals who may not have "the demeanor that is expected in a governmental, bureaucratic setting where efficiency is highly valued." Instead of blocking out a 15-minute period for a meeting, the leader might say, "If this is important, let's spend a few days on it." To effectively engage a community, data users may have to ignore how management gurus say a meeting should be run; instead, "just follow your grandmother's advice: sometimes you just need to listen and not say anything."

Source:

Testimony of Dr. Phillip Smith, Indian Health Service Institutional Review Board, NCVHS Subcommittee on Privacy, Confidentiality and Security, April 17, 2012.

A Refugee Community's Expectations

One community health promotion project found that people in some immigrant and refugee communities did not expect privacy and did not understand how sharing information might cause harm. In the same project, researchers encountered a clash between the U.S. emphasis on individuals and some communities' emphasis on the family unit. Families did not want the "head of household" representing the family on a survey; instead, they wanted the family to complete the survey as a unit. Although the organization's Institutional Review Board found this approach disturbing because it would not preserve confidentiality among family members, the board agreed to proceed by following the preferences of people in the community.

Source:

Linda Silka, Ph.D., University of Maine (interview and correspondence)

Taking Neighborhood Health to Heart

Taking Neighborhood Health to Heart (TNH2H) started as a community-based participatory research project involving diverse urban neighborhoods in Denver, the University of Colorado Denver, and the Stapleton Foundation. Funding from the National Institutes of Health allowed TNH2H to study the impact of the built and social environment on health and health disparities among neighborhood residents. Information about the project is available at TNH2H.org.

TNH2H involves community members at every stage. In addition to involving them in creating the survey, community members gave information that helped develop the surveys, and people in the community were employed to administer surveys. The outcomes of the original research project were shared with neighborhoods. In addition, the community identified and directed follow-up studies and outcome dissemination.

Laws and regulations do not routinely require the level of involvement from community members in research that is found in TNH2H. By going beyond legal requirements of openness, transparency, and choice, TNH2H earned the trust of the community and has successfully engaged the community in

improving the health of its members.

Source:

Debbi Main, Ph.D., University of Colorado Denver (interview and correspondence)

PINE Study

The PINE Study is the joint product of collaboration among the Chinese Health, Aging, and Policy Program at Rush University, Northwestern University, and more than 20 community services organizations in the Chicago area, including the Chinese American Service League and Xilin Asian Community Center as the main community partners. This academic–community partnership is guided by a community-based participatory research approach. The PINE Study was designed to identify actionable health policy concerns among a population of individuals whose preferences and service needs are poorly understood. Older Chinese adults are hard to reach because they tend to distrust programs run by the federal government, due to the Chinese community's past experience with harsh violence and decimation. The issue is further compounded by vast cultural and linguistic barriers.

During 2011–2013, the PINE Study carried out face-to-face interviews with 3,159 community-dwelling older adults from 60 to 105 years old. The multilingual staff interviewed participants based on their preferred languages and dialects, including English, Cantonese, Taishanese, Mandarin, or Teochew dialects. Data were collected using Web-based software that recorded simultaneously in English, Chinese traditional, and simplified characters. Due to the careful planning and community engagement, the response rate was 91%.

The result of the effort was The PINE Report, a comprehensive study that examined the health and well-being of Chinese older adults in the greater Chicago area—the largest cohort of older Chinese adults ever assembled for epidemiological research in Western countries. The report revealed that individuals in this population are affected by medical comorbidities, physical disabilities, low health care utilization rates, psychological distress, social isolation, and elder abuse at higher rates than the average older adult in the United States. Many experience low acculturation levels, financial hardship, and insufficient social support. The PINE Report identified opportunities for family members, community stakeholders, health professionals, and policymakers to improve the health and well-being of older Chinese adults.

Source:

Dong XQ, Chang ES, Wong E, Wong B, Skarupski KA, Simon MA. Assessing the health needs of Chinese older adults: Findings from a community-based participatory research study in Chicago's Chinatown. J Aging Res 2011 2010:1–12.

MyHealth Access

MyHealth Access Network is a nonprofit coalition of more than 200 organizations in northeastern Oklahoma, with a goal of improving health care quality and the health of area residents while controlling costs. The organization was chartered to facilitate communications and connections among participants in the health care systems. MyHealth does not directly provide care, but gives those that do the technology, information, communications, and analytics to support improved care quality and reduced costs (see http://www.myhealthaccessnetwork.net/).

MyHealth Access Network engaged the community in a 100-day planning process that involved 200–300 people. At the beginning, participants agreed to focus on the objectives of health improvement and quality. They recognized that a primary focus on privacy and security, without starting by defining the return on investment, would scuttle any effort to share and use health data to improve health.

A subset of task forces was formed to address specific issues, including content, clinical, privacy and security, and costs. The recommendations and findings from these groups were reviewed by top-level governance to create a plan.

Throughout the process, facilitators refused to allow conflict to become disengagement, which led to the model's widely recognized success.

Source:

Interview with David Kendrick, M.D., M.P.H., MyHealth Access

Research on biological samples from members of the Havasupai Tribe

Members of the Havasupai tribe gave DNA samples to Arizona State University (ASU) researchers in the early 1990s. The researchers suggested that the DNA samples might provide information about the tribe's very high diabetes rates. In the early 2000s, however, a tribal member heard a presentation about the data that addressed migration, mental health, and "inbreeding."

The tribe was deeply disturbed that biological samples taken to

assist tribal members with a specific health concern were used in ways that directly challenged beliefs of tribal members, while also stigmatizing all people of the tribe. This shows that harm is not only caused when personal health data are disclosed (as in the hospital discharge data set), but when every person in a small group can be stigmatized.

After a lawsuit was filed, ASU agreed to a settlement to "right the wrong" in using the data in a way that violated tribal members' right to consent.

Source:

American Indian and Alaska Native Genetics Resource Center website: http://genetics.ncai.org/case-study/havasupai-Tribe.cfm.

Appendix D: Worksheet and Checklists

Purpose Specification Worksheet

Accountable entity or individual(s) _____

Describe the purpose of data use

Describe the role of the community and affected individuals in specifying the purpose of data collection or use

Describe data elements needed to achieve the purpose

From what source(s) will you get the data?

- ☐ Federal public data sets
- ☐ State public data sets
- ☐ Medical records
- ☐ Original survey
- ☐ Other

Will data be repurposed?

- ○ Yes ○ No

What potential adverse consequences, if any, do you anticipate:

- ☐ Risk of breaching individual's privacy or confidentiality
- ☐ Negative impact on community
- ☐ Stigmatization of individuals or small groups

Describe plans to lessen possible adverse consequences (e.g., notice, data protection, community consultation)

Describe possible future use/repurposing

Describe procedures for considering, and limits on, unplanned use

Describe how to evaluate the need to consider additional consent when repurposing data

Data Quality and Integrity Checklist

Data Collection

Accountable individual/entity:_____

Describe the plan for community engagement in the data collection process.

Are either original or repurposed data collected following acceptable data collection and use practices?

- ☐ If the organization lacks expertise in data collection best practices, look for outside assistance from a researcher, health care provider, state health department, or other organization with expertise in data collection and entry
- ☐ Sample is representative of population of interest
- ☐ Data collection procedures set up and documented before data collection
- ☐ Training for those engaged in data collection
- ☐ Require those collecting data to sign confidentiality agreements
- ☐ Audit data collection processes
- ☐ Training for those entering data (if a separate process)
- ☐ Audit data entry processes

Repurposed Data

- ☐ Data source is trustworthy

Merging Data Sets

Accountable individual/entity: _____

- ☐ Are the populations the same for the different data collection efforts?
- ☐ Do survey questions and response categories match?
- ☐ Might differences in survey administration dates affect survey results?
- ☐ What were the survey sample designs?

Describe methods to be used when merging data sets.

Data Analysis

Accountable individual/entity: _____

Describe valid methods for analyzing qualitative or quantitative data, or identify the individual or entity that will do the analysis

Reporting Results

Accountable individual/entity: _____

Describe how reported results will protect communities, subgroups, or individuals from bias or stigma.

Describe protections to ensure accurate reporting of results.

Data Security

Accountable individual/entity: _____

- ☐ Identify ways to protect data integrity/security
- ☐ Encrypt personally identifiable information on mobile devices
- ☐ Create a de-identified data set
- ☐ Use valid methods if producing a de-identified data set
- ☐ Limit password-protected access to identifiable data to those with a need to know
- ☐ Limit the ability to delete, add, or change data to those with appropriate training and need
- ☐ Store paper records with identifiable information in a different place from records that do not contain identifiers

Openness, Transparency, and Choice

Accountable entity or individual(s): _____

Describe community engagement in the data collection process

Determine the appropriate level of disclosure

- ☐ Community notice (describe)

- ☐ Small group notice (describe)

- ☐ Individual notice (describe)

- ☐ Create a feedback loop with participants/community to report findings and recommendations (describe)

Data Use Agreement Checklist

Data use agreements designed to limit re-identification of de-identified data should, at a minimum, address the following elements:

- ☐ Define the scope of data use
- ☐ Require recipient to use safeguards to prevent use or disclosure not allowed in the scope of the agreement
- ☐ Require recipient to report to the data source any use or disclosure not allowed in the scope of the agreement
- ☐ Require recipient's agents, such as subcontractors, that receive the data to agree to the same restrictions and conditions that apply to the recipient
- ☐ Require the recipient to agree to refrain from identifying or contacting individuals whose health information is contained in the shared data set.
- ☐ Define scheduled monitoring by data source and/or assurances by data recipient confirming that terms of the agreement are being honored
- ☐ Specify consequences of the data recipient's failure to comply with terms of the agreement
- ☐ Specify who bears the cost of enforcing the agreement if the data recipient is alleged to violate the agreement

If you are being asked to sign a data use agreement in order to receive data, find out:

- ☐ What laws or regulations, if any, govern the data sharing and what the laws or regulations require of you as a recipient of data
- ☐ What the document allows you to do and not do with the data
  - ☐ How does the document define the scope of use?
  - ☐ Limits on attempts to re-identify or contact individuals associated with the data
  - ☐ Who can see or work with the data, inside or outside of the organization
- ☐ Can you provide physical or technological safeguards that must be in place to secure the data under the agreement?
- ☐ Can you meet requirements to audit data use or track access to data?
- ☐ What are your duties if there is a breach of the agreement
  - ☐ Reporting? To whom?_____
- ☐ How will you address any allegation that you or your agents have breached the agreement?

Limited Data Set Checklist

If receiving a limited data set (LDS) from a covered entity, an organization should confirm that the data use agreement includes the following elements[1]:

- ☐ Identifies the receiving organization as the recipient of the LDS
- ☐ States that the LDS will be used only for research, public health, or health care operations
- ☐ Describes the purpose for using the LDS
- ☐ LDS recipient agrees to refrain from using or disclosing the LDS for any purpose not specified in the agreement
- ☐ LDS recipient agrees to use appropriate safeguards to prevent use or disclosure not specified in the data use agreement
- ☐ LDS recipient agrees to report LDS use or disclosure of LDS not specified in the DUA
- ☐ LDS recipient agrees that its agents, such as subcontractors, that receive the LDS agree to the same restrictions and conditions that apply to the LDS recipient
- ☐ LDS recipient agrees to refrain from identifying or contacting individuals whose health information is contained in the LDS

1 The University of Wisconsin has compiled "HIPAA Privacy and Security Rule Policies and Procedures," including limited data set information as noted above. The compilation is available from: http://hipaa.wisc.edu/hipaa-policies.htm.

NCVHS Membership, September 2014

Larry A. Green, M.D., Chair

John J. Burke, M.B.A, MSPharm.*

Raj Chanderraj, M.D., F.A.C.C.

Bruce B. Cohen, Ph.D.

Llewellyn J. Cornelius, Ph.D.

Leslie Pickering Francis, J.D., Ph.D. *Subcommittee Co-Chair

Alexandra Goss

Linda L. Kloss, M.A., RHIA *Subcommittee Co-Chair

Vickie M. Mays, Ph.D., M.S.P.H.*

Sallie Milam, J.D., CIPP, CIPP/G*

Len Nichols, Ph.D.

W. Ob Soonthornsima

William W. Stead, M.D.

Walter G. Suarez, M.D., M.P.H.*

James M. Walker, M.D., FACP

* Member of the Privacy, Confidentiality and Security Subcommittee

Lead Staff for the Subcommittee
Maya A. Bernstein, J.D. ASPE

Executive Staff Director
James Scanlon
Deputy Assistant Secretary,
Office of Science and Data Policy
Office of the Assistant Secretary for Planning and Evaluation,
DHHS, ASPE

Acting Executive Secretary
Debbie M. Jackson, M.A.
Senior Program Analyst
Classifications and Public Health Data Standards Staff,
Office of the Director
National Center for Health Statistics, CDC

This report was written by NCVHS Consultant Writer Maureen Henry, in collaboration with NCVHS members and staff.

The National Committee on Vital and Health Statistics (NCVHS) is the statutory [42 U.S.C. 242k(k)] public advisory body to the Secretary of Health and Human Services (HHS) for health data and statistics. The Committee provides advice and assistance to the Department and serves as a forum for interaction with interested private-sector groups on a variety of key health data issues. The Committee is composed of 18 members from the private sector who have distinguished themselves in the fields of 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. Sixteen of these members are appointed by the HHS Secretary to terms of four years each, with about four new members appointed each year. Two additional members are selected by Congress.

For more information, see the NCVHS website:
http://www.ncvhs.hhs.gov/

Centers for Disease Control and Prevention
National Center for Health Statistics