NCVHS

Toolkit for Communities Using Health Data

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

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 advance health at the individual, subgroup, and community level have revealed the need for guidance on the meaning and application of data stewardship for these users. Participants in these efforts have focused on the needs of community-level organizations. NCVHS chose to create the *Toolkit for Communities Using Health Data* to provide a substantive introduction to the elements of data stewardship to communities seeking to use data.

For the purpose of this document, "communities" are deliberately 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 Denver, 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 process.

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 healthcare systems. Now communities are able to 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 community health data users to promote community and individual health. If used effectively, 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, or
- Opportunities to improve the health of the community as a whole and 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 seeking to use their data. In this Toolkit, we have attempted to pull together important themes in stewardship—proper data protection and use—and, where relevant, to refer users of community health data to some of these resources.

Introduction

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.

The purpose of this Toolkit is to support communities that are using data by promoting sound stewardship practices, while helping them to avoid the missteps and potential harm that can result when data users fail to 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 a 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 conclusions 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 level.

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 specific tips for data users. Detailed descriptions of stewardship principles are provided, along with check-lists 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 & 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 sections 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.

---

[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 Sibelius, "A Stewardship Framework for the Use of Community Health Data," (Dec. 5, 2012), *at* 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
- Check-Lists
- Data Use Agreement Template

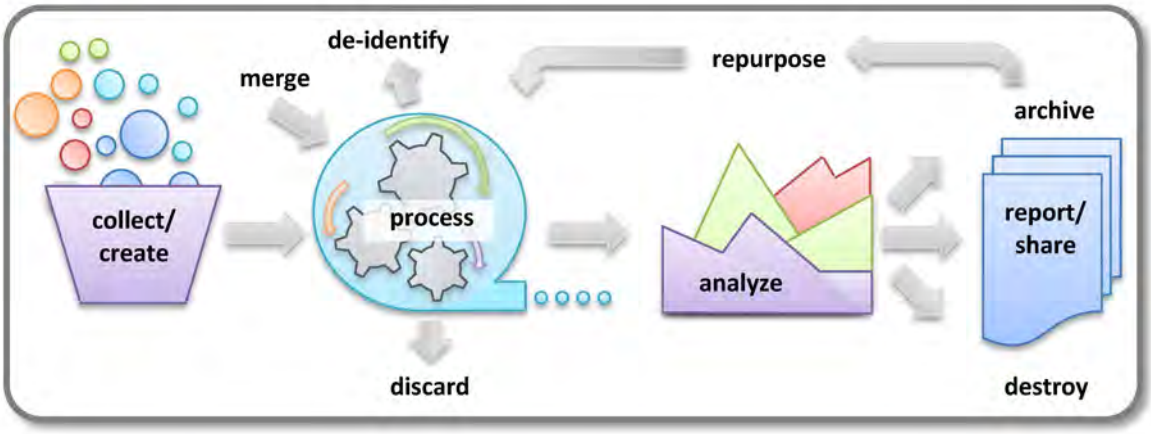
Data Lifecycle

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 advance health must undertake that are outside of the data lifecycle, such as conducting a literature review to understand the current knowledge on the topic and to better frame the purpose of the inquiry.

Data Life Cycle



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.

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

Until recently, the data in patient medical records were used primarily for patient care, payment, and the operations of healthcare institutions. Data abstracted from paper medical records were used for research and other purposes but it was costly and difficult to extract data manually. 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 physician visit that is entered into the history and physical portion 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 extensively 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 so doing, they might employ 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 significantly changed how data are managed at all lifecycle stages from origination to eventual destruction or archive. Technology speeds the capture of data and when it is available for use. It can help to maintain a description of the characteristics of data—what are called "meta-data"—including who collected the data, when it was collected, what permissions or restrictions attach to it, flaws or limitations of the data, and other such characteristics. Technology can also be used to establish 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 volumes of electronic data,
- Process and analyze large data sets efficiently,
- Enrich data sets by merging data from different sources,
- Repurpose data in ways not conceived when the data were collected
- Access data remotely, and
- Copy or transmit data rapidly.

For example, 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 advance 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. Opportunities include:

- Understanding health at a granular level, such as geo mapping health data to provide an understanding of how disease affects individuals living on a particular block within a community
- Evaluating the impact of programs on health by linking data about who received an intervention with data from a community-wide health information exchange and claims data

But with opportunity comes risk:

- Data breaches are evidence that data security is challenging, even for large companies and governments with substantial resources.
- Data elements that appear to be the same may have different meaning across systems impeding accurate interpretation.
- Repurposing, while an opportunity, can cause harm when it occurs without appropriately engaging individuals and communities, as shown in several of the Case Studies described later in this Toolkit.
- Problematic inferences due to the analysis of electronically processed data may result in social stigma and harmful reputational effects for the wrongly categorized individuals.

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

Governmental and Non-Governmental Data Collectors and Users

Data stewardship for non-governmental data collectors or users has much in common with, but is not identical to, data stewardship for governmental data collectors or users. Nevertheless, both government and non-government data stewards must act in accordance with 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 obligation to serve members of the public by openly and transparently sharing data. Non-governmental data users and collectors do not share that affirmative obligation, although sharing data to serve the community may be consistent with stewardship principles.

Data Stewardship

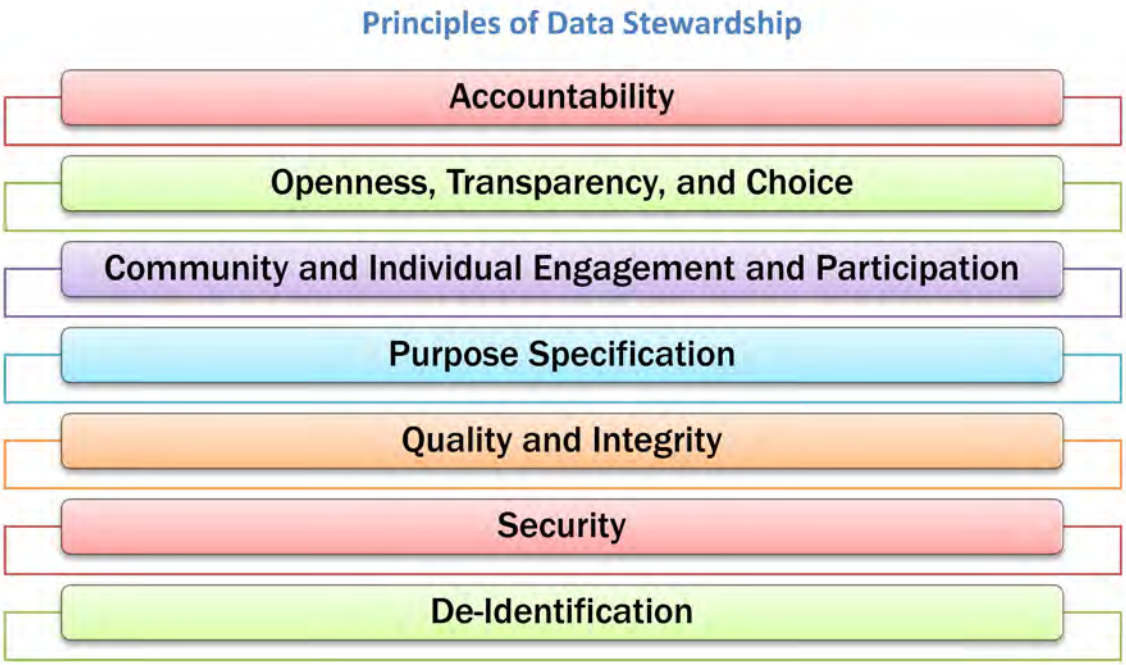
Data stewardship is a responsibility, guided by principles and practices, to ensure the knowledgeable and appropriate use of data. More specifically, stewardship of health data recognizes the benefits to society of using personal health information to improve understanding of health and health care while 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 elements.

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

Non-Linear, Overlapping Concepts

The figure showing the elements of data stewardship below suggests that stewardship elements are discrete and linear. On the contrary, as is acknowledged throughout the Toolkit, elements overlap, and the stewardship process may require data users to loop back or jump forward as circumstances demand.



Accountability

The first thing a community should do when considering a new data analysis project is to assign responsibility for accountability for all aspects of the project. Accountability means that an individual or entity has formal responsibility for

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

One person might be accountable for every element of data stewardship across the data lifecycle, or different people or entities might be accountable for different parts of the process. It is important, however, to assure that data users can identify the accountable person. Also, when a failure of accountability occurs, the accountable individual or entity should face consequences, and the responsible entity should provide remediation to individuals whose data were compromised.

Data users should identify who is accountable at each step of the data lifecycle to assure that the elements of data stewardship are honored—from project conceptualization, through initial collection and use, to data destruction, storage, or repurposing. The responsibilities might be divided among different parts of the lifecycle or according to the different stewardship elements.

Failure to identify and address concerns regarding proper stewardship may lead to a variety of 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 a useful tool 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 data use agreement if a data user violates the agreement.

When Data Users Are Asked to Sign a DUA

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 items outlined on the check list at the end of this section. 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 in an individual or entity
- Different people may be accountable for different phases of the data lifecycle or different stewardship elements

Accountability

- Accountable individual or entity should be named and held responsible for stewardship
- DUAs are one way to establish accountability among data users

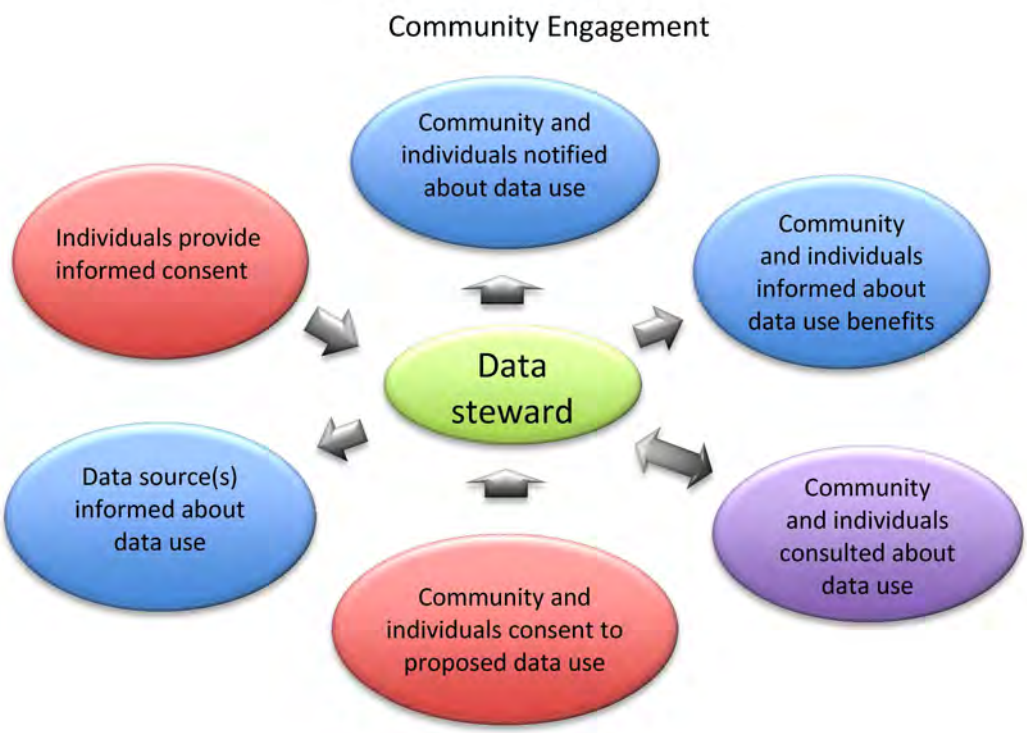
Consider This: Accountability Ombudsman

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 with which Vanderbilt collaborated needed a single person who could help them navigate the network of accountability. The eMERGE network at Vanderbilt identified an ombudsman: an assigned, responsible person who could explain the organization's accountability policies and procedures to members of the community and who could assure that community members' 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. Failure on the part of data users to maintain openness and transparency, and failure to offer choices to individuals and communities when required or appropriate, can create unwelcome surprises, destroy trust, and produce serious adverse effects on the ability to use health data to advance health. The Toolkit provides 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 trust.



Community engagement alone may not, however, be adequate to assure openness, transparency, and choice in cases where individuals' preferences may not be aligned with 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.

Openness, Transparency, and Choice

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 have the opportunity to ask questions about, comment on, or object to data use. Data users may also be required to provide sources of data, such as health care providers, public health agencies, or researchers, with similar information.

Individual notice

Individual notice may be warranted when those whose data are being used are identifiable, for example, by name or residential 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 intend to use protected, personally identifiable data without other prior notice, they may need to provide individual notice. In some instances law 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, email, 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.

Data users should use caution when the notification itself has the potential to 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 Human Immunodeficiency Virus (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 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. Appendix F provides a template for organizations that wish to use a Notice of Privacy Practices.

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 greater detail below.

Community notice

In some instances, notice is provided to the community, not individuals. Different techniques may be used to provide notice to a community, including:

- Community meetings or town halls
- Booths at community events
- Flyers or notices available at libraries, community centers, government offices
- Web sites 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 warranted. For example, if data use were to affect Asian women with cancer, notice may be provided in a newsletter targeted at this population, shared on blogs frequented by these community members, 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.

Consider This: Engaging individuals and communities preserves the use of fetal blood spots to advance 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 web site 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.

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 challenges 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 cost of notice are very high or providing notice is impracticable, 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 provided broadly to communities or subgroups within a community, or targeted to the individuals whose data will be used.

Consider This: 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.

Consent

In addition to notice, individuals may have the opportunity to choose whether their data may be used. 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] 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 are outside of the scope of any original consent. For example, individuals who have consented to the use of their data to study diabetes might need to be offered the explicit opportunity to choose whether they wanted to participate in a study of correlations with mental illness or substance abuse. Even if data use is not governed by laws or regulations, however, a data user should evaluate 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 prior to 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 exceed the benefits.

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 in lieu of obtaining individual consent. This type of consent is appropriate where the risks to community members are relatively low. It may not be appropriate when risks to individuals or small subsets of individuals in the community are high.

Opt-In/Opt-Out

In some cases, individuals may be offered the choice between allowing their data to be used or not used. Opt-in and opt-out provisions typically have a default. With an opt-in approach, individuals must take action to have their data included in 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.

---

[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 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 data
- When using de-identified data
- When using aggregated data supplied by governmental agencies or data aggregators
- 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 illustrated in many case studies.

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

Cautionary Tale: Repurposed Use of Blood Samples

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

Mechanisms for engaging community members

Data users can engage community members through a number of different mechanisms. When determining community engagement methods, data users should consider which mechanisms may provide legitimacy for the data effort. In a politically polarized community, for example, elected officials may not be perceived as representing the interests of many voters. The following briefly summarizes some approaches to community engagement.

Community Leaders

Community leaders can sometimes serve as a proxy 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. It may be convenient to use community leaders as proxies for individuals, but 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 mechanism for engaging communities. Focus groups have the potential to generate a deep understanding of how individuals feel about an issue. There are guidelines available on how to effectively conduct a focus group available at http://assessment.aas.duke.edu/documents/How_to_Conduct_a_Focus_Group.pdf. Like engagement through community leaders, focus groups can miss issues relevant to subgroups if members of subgroups are not represented among 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 assembling 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 at 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.

Consider This: The Community Takes the Lead

In the initiatives undertaken by Taking Neighborhood Health to Heart (a community group in Denver concerned about local rates of heart disease), the community shares responsibility in guiding the questions to be asked, research to be conducted, and release of data. In some cases, community members are hired to collect survey data. Because the community is an active participant in all phases of research, the initiative has successfully learned about issues that might never have been addressed for fear that results would be used to stigmatize community members.

Community and Individual Engagement

Opportunities for engaging community members across the data lifespan

Purpose Specification

When conceptualizing 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 those seeking to use their data to engage communities:

- Community members can administer surveys, potentially improving participation and response rates (see Taking Neighborhood Health to Heart in the accompanying box, and in 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 those using their data to 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 lack a nuanced understanding of community dynamics. Communities can be very helpful in reviewing findings and interpretations of findings prior to releasing findings 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, 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 identify areas of sensitivity or concern and open

communication channels to address concerns. Data users from outside the community may not anticipate how the data may adversely affect communities. Studies of prevalence of health challenges such as sexually transmitted diseases, substance abuse, behavioral health, or genetic disorders, whether derived 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 in order not to stigmatize groups or individuals.

Consider This: Stigmatization in the AIDS epidemic

In the initial days of the HIV 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* (supp) 45: 461s9-4620s (1985). The result was widespread fear of Haitian immigrants and a significant drop in tourism to Haiti. One of the physicians 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: Oxford University Press, 2000, pp. 28-29.

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 wish to learn from, but perhaps not publish, results.

Consider This: Engaging a distinct community subgroup

The *P*opulation Study of Ch*I*Nese *E*lderly (PINE) identified actionable concerns among older Chinese adults in Chicago, a community cohort that was less well understood. By engaging over twenty community groups and by using multilingual staff to interview participants according to their preferred language and dialects, the survey response rate was 91%. The result of the effort was reported in The PINE Report, which revealed 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 the average 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 wellbeing of older Chinese adults.

Community and Individual Engagement

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 sub-groups 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

Consider This: Engaging the Community in Health Information Exchange

Health Information exchanges allow 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 at risk has decreased information sharing through exchanges, likely to the detriment 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 outset, 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.

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 be required to explain the purpose for the data use. Even for data that is publicly available, articulating the purpose is important if the data use is to achieve its intended goal.

Purpose specification has several 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 adverse 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 typically driven by a question that the data may help to answer.

Consider This: 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 four 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.

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 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 request consent in the primary study for samples to be used in later studies defined in the consent.

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
- future uses will surprise individuals or communities with unexpected, perhaps even unwanted, results.

Repurposed Data

Repurposed data are collected for one purpose 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 demonstrate, for example, 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 demonstrates how data users can more likely avoid problems encountered by data users who failed to consider the risks of repurposing.

Cautionary Tale: Repurposing data without individual or community engagement

The vast majority of 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 over five million biological samples.

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 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. In other cases, state laws or regulations allow the sharing of government health data only for specific purposes.

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 original purpose may trigger the need for additional notice or consent.

---

[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 for the broad-based healthy, sustainable, and livable communities' movement. Registered users have free access to:

- Thousands of map-able geographic information systems (GIS) data layers and tables displayed at varying geographies for all communities in the United States;
- An application program interface (API), providing free interoperable access to data;
- Contextualized mapping, reporting, data visualization, and sharing abilities;
- Searchable profiles of place-based community initiatives and multi-sector collaborations;
- Peer learning opportunities to explore similar topics and share best practices;
- Spaces for individuals and communities to share narratives, interviews, videos, images, documents and other online resources;
- Searchable profiles of hundreds of place-based community initiatives and multi-sector collaborations working towards healthy/sustainable/livable/equitable communities; and
- Peer learning opportunities with colleagues across the country exploring similar interests and challenges."

See "About" *available at* www.communitycommons.org.

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

Consider This: 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 whether 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 prevailed 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 wish to try this option if they encounter confidentiality concerns. *Southern Illinoisan v.Ill. Dep't Of Pub. Health*, 218 Ill. 2d 390 2006).

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.

Data 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 posed 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.

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.

Data Quality and Integrity through the Lifecycle

Review the Literature

Data users should research and evaluate what has already been done; doing so helps to assure the quality of data. This 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 the duplication of effort, and avoid the past mistakes of others.
Trustworthiness of Data SourceTo assure the quality of data, users should assure that original data are collected in accordance with generally accepted procedures and that sources of repurposed data are trustworthy. A trustworthy data source would have the ability, for example, to provide the user of repurposed data with assurances about how data were collected, entered into a database, and stored. The Quality and Integrity Check-list in Appendix D enumerates steps for users to consider.

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 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, when doing so, 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

- Results of the a survey of nutritional habits of adolescents administered by different school districts in different cities in a state could be combined to increase the statistical power of the study
- 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 assure that a merger or aggregation is valid and that the data retain integrity. In determining validity, data users should ask the following questions:

- 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?

Answering these questions may require expert advice. Substantive issues about combining data should be resolved before any statistical consultation can take place.[4]

Summary

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

---

[4] For a detailed discussion about how to evaluate the validity and integrity of merging data sets *see* U.S. Dep't of Health & Human Servs., Ass't Sec'y for Planning & Eval., *Data on Health and Well-being of American Indians, Alaska Natives, and Other Native Americans: Data Catalog*, Contract No. 233-02-0087, App. B: Data Set Aggregation, B-1 (Dec. 2006), at http://aspe.hhs.gov/hsp/06/catalog-ai-an-na/report.pdf.

Security

Securing data means protecting the confidentiality, integrity, and availability of the data. In other words, 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 the door of an office 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. This applies to any information that might identify individuals (whether particularly intimate or not) such as their names, email address, account numbers, health information or many other types of information.

Some of the primary threats to loss of data include using weak passwords, failure to back-up electronic data, infection by viruses or malware, 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 the following steps:

- Evaluate anticipated risks
- Develop a plan to mitigate 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, so that data cannot be accessed if the computer is lost or stolen
- Automated back-up processes to protect against accidental data loss
- Training for employees on security measures

Like the question of notice, data users must consider the need to secure data, and the costs of doing so, in a way commensurate with the risk of data loss, inappropriate access, or manipulation.

Examples of mitigation techniques to support data security

- Physical
    - Install locks on cabinets or rooms where paper records are stored
    - Maintain 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
    - Conduct a risk analysis

  - Establish policies and procedures for accessing paper records, for disposing of data, or for adding new equipment on a network
  - Train those with access to sensitive information about data security
  - Require robust passwords
  - Control who has access to view or change the data
  - Conduct due diligence on employees handling data
  - Implement an incident response program
- Technical
  - Maintain logs of system access and exfiltration of data
  - Encryption
    - Specific elements in a data set
    - Data set as a whole
    - Devices that permit access to the data set, such as laptop computers
  - Implementation of monitoring to scan for and identify cyber-attacks

For more detailed information about security, the National Institute of Standards and Technology publishes useful guides to assessing and maintaining data security which may be useful to organizations even if they are not federal agencies.[5] The Office for Civil Rights of the United States Department of Health and Human Services also publishes security guidance in plain language for entities covered by the HIPAA Security Rule which may be useful to organizations even if they are not covered by that rule.[6]

The 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.

---

[5] The National Institute of Science and Technology's Computer Security Resource Center publishes guides on a variety of topics. A list may be found at http://csrc.nist.gov/publications/PubsSPs.html.

[6] The Office for Civil Rights publishes educational materials to help covered entities learn more about the HIPAA Security Rule and other sources of standards for safeguarding electronic protected health information that may also be useful to entities that are not required to comply with the rule. In particular, OCR published the *HIPAA Security Information Series*, a group of educational papers which are designed to give HIPAA covered entities insight into the Security Rule and assistance with implementation of the security standards. These are available at http://www.hhs.gov/ocr/privacy/hipaa/administrative/securityrule/securityruleguidance.html. We recommend readers start with *HIPAA Security Series 1: Security 101 for Covered Entities*, for an overview of basic concepts. Data users may also wish to consult *HIPAA Security Series 7: Security Standards: Implementation for the Small Provider* which describes basic topics. These other resources are also available: Privacy and Security Training Games: http://www.healthit.gov/providers-professionals/privacy-security-training-games; Guide to Privacy and Security of 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; 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 refers to 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 the identity of individuals and information about them. The promise of data de-identification is that, by removing directly identifying elements and otherwise manipulating data, 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 sensitivity of the data. 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. The purpose of data de-identification is to make it very difficult to link data to a specific individual, permitting the study of a variety of sensitive issues while significantly reducing the risk of disclosing personal or confidential information. Aside from organizations that must follow the HIPAA de-identification methods, there is no standard, universally adopted de-identification method that is used throughout health care.

Identity v. Attribute Disclosures

There are two areas of concern regarding re-identification. The first is called identity disclosure. Identity disclosure happens when an outside party is able to assign an identity to a record in a disclosed dataset.

The second type is called attribute disclosure. Attribute disclosure allows an outside party to attribute characteristics to someone in the data set even if they have not been individually identified. This form of disclosure is of primary concern in summary data releases. 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 permit 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 particular category of diabetes patients defined by race and sex had good control of their blood glucose levels.

If the opposite is true, a cell has 100% of a particular subgroup in a sample showing a particular 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.

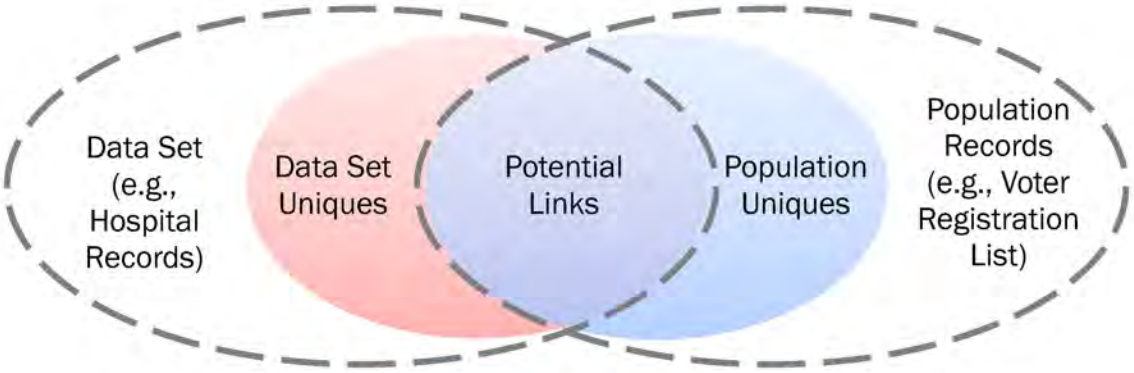
Simple De-identification

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. To understand why, it is necessary to understand how re-identification attacks are performed.

Even when an administrator removes all of the data fields he or she thinks might be uniquely identifiable from a data set, it is possible for an attacker to unlock the identity of the subject of the record by discovering pockets of uniqueness remaining in the data. This sort of re-identification is possible because, even absent a specific identifier, certain combinations of values may be so rare as to serve as a "fingerprint" which could only point to one person. A re-identification attack attempts to locate the unique fingerprints in a de-identified dataset,

and then search for that same unique fingerprint in another dataset which does contain unique identifiers. This technique can be aptly illustrated using a Venn diagram:

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 dataset without phone numbers removed. In a more complex form, this type of 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 that there is only one such individual in that zip code of that sex who was born on that day. This external linkage through uniqueness is the risk that de-identification attempts to protect against.

Re-identification Using Public Records[8]

2014 Disease Combined STI Reporting			
Age	Zip Code	Gender	Diagnosis
27	25314	Male	Chlamydia
18	25311	Male	HIV
45	26003	Female	HPV
82	24901	Male	HIV

2014 Voter Registration Records			
Birthdate	Zip Code	Gender	Name
09/05/1987	25314	Male	Victor Richardson
12/04/1995	25311	Male	John Doe
03/18/1969	26003	Female	Jane Doe
01/31/1932	24901	Male	Josiah Doe

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

---

[7] *Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability (HIPAA) Privacy Rule*, Department of Health and Human Services: Understanding HIPAA, *available at* http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html, 1 (last visited Sept. 24, 2014).

[8] *Id.*

Individual Level De-Identification

Data users can de-identify individual records through a number of techniques. Those most commonly employed are suppression, generalization, and distortion. Suppression occurs when information is completely removed from the data set.

Example Dataset

Age (Years)	Gender	ZIP Code	Diagnosis
15	M	00000	Diabetes
21	F	00001	Influenza
36	M	10000	Broken Arm
91	F	10001	Acid Reflux

Direct identifiers such as names and social security numbers are common examples of individual data that is completely suppressed. Some data such as birthdates and zip codes, however, cannot be completely suppressed without destroying the utility of the dataset.

Example Dataset - 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. Generalization occurs when a particular variable, such as age, is divided into broader categories such as five year age spans. Generalization is often extremely effective at balancing utility and privacy in a data disclosure.

Example Dataset - Generalized

Age (Years)	Gender	ZIP Code	Diagnosis
< 21	M	0000*	Diabetes
21 ≤ 34	F	0000*	Influenza
35 ≤ 44	M	1000*	Broken Arm
>45	F	1000*	Acid Reflux

Distortion may also be used to de-identify data, but, with regard to health data, distortion often destroys the reliability with which the data can be used to draw effective conclusions.

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

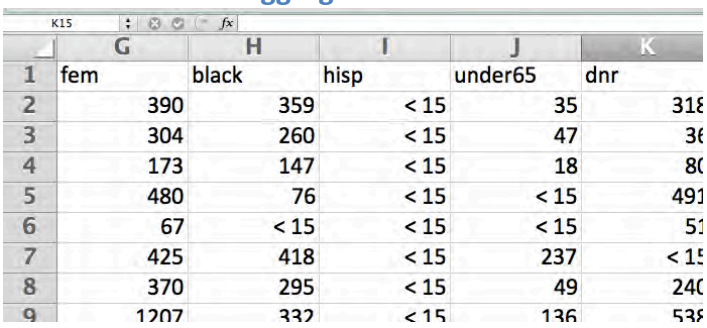
de-identification can be particularly effective at maintaining utility while protecting the confidentiality of the data. There remains a risk, however, of inadvertent attribute disclosure. For example, the table below logically implies that all Hispanic females enrolled in the Healthyville School District 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 Males	85	40	15
White Females	90	12	7
Black Males	45	15	8
Black Females	50	11	13
Hispanic Males	10	5	7
Hispanic Females	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 particular type of cancer, those four patients are at high risk of being identified. In the illustration of aggregated data in the figure that follows, the number of individuals 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 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 could seek expert advice for assistance in methods to further mitigate these risks.

In addition, data users can employ data use agreements, discussed below, to limit attempts at re-identification. Another approach is to ask individuals whose data are used if they would consent to data use even if there were a risk of re-identification.

While 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 health.

Quantifying and Evaluating the Risk of Re-Identification

Evaluating risk of re-identification can be a very technical process that requires substantial expertise, but there are general principles that users of community health data can follow 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.

Cautionary Tale: Small Cell Sizes

An academic used state vital records data from death certificates to investigate cause of death from a variety of causes. This researcher was able to identify a single individual because of small cell size. As a consequence, the government agency that supplied the data decided to increase the suppression criteria from 5 to 10. They implemented 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 one conducts an analysis for which any of the cell sizes are less than 10, the cell will come up blank or just indicate "<10".

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 the analysis of the combined data sets might 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 evaluation of the risks of re-identification for new mergers of de-identified data sets and for any and 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 on how HIPAA-covered entities can evaluate risk of re-identification,[9] and that guidance may also be useful to entities not covered by HIPAA, 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 DUAs when researchers use what are called "limited data sets," data sets that have been created from protected health information by removing all identifiers except certain information about dates and locations. Users obtaining de-

[9] Office for Civil Rights, U.S. Dept. of Health & Human Servs., Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule (Nov. 26, 2012), *available at* <http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html>

identified data sets also may be required to enter into a DUA with the entity supplying the data to promise to protect the data against re-identification or to make additional privacy and security arrangements. 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.

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.

The State of Washington learned from the experience and put in place a system using data use agreements that, among other things, researchers accessing the data agreed that they would not try to re-identify individuals in the de-identified data set.

Summary

- De-identification can be used to limit the risk that individuals' confidential or private data will be disclosed.
- Two types of de-identification include:
  - 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.

Appendix A: Definitions

The following definitions explain how terms are being used within the context of the Toolkit, although the definitions are never radically different from other common uses.

Community

A community is an interdependent group of people who share a set of characteristics and are joined over time by a sense that what happens to one member affects many or all of the others.[10]

Confidentiality

The treatment of information that a person has disclosed in a relationship of expected 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 numbers, 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 five years old with asthma living in a specific zip code, or about 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 grants certain rights to individuals—for example, to obtain copies of their medical records—and imposes obligations on healthcare 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," a federal regulation for the Protection of Human Subjects in Research, to assure that research involving people meets legal and ethical requirements. Federal and state laws and regulation dictate what research must be approved by an IRB.

---

[10] This is the definition of community used in the report of the National Committee on Vital and Health Statistics, *The Community as a Learning System: Using Local Data to Improve Local Health*, p. 8 (Dec 2011).

Appendix A: Definitions

Notice

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

Protected Health Information or PHI

A term of art referring 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 advance the health of community, or 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 familiarize themselves to 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 subject research across a range of settings, including research conducted by universities, state and local governments, and non-profit 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 obligations and limitations of the entities from which communities seek data.

This Toolkit does not give data users everything they need to know about HIPAA or human subject protection rules and regulations. Rather the goal is to alert data users to circumstances when they need to seek further guidance from attorneys or compliance experts to assure 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, it should consider the types of protections and inquiry required of data protection and sharing in the HIPAA Privacy Rule and human subject protection laws and regulations. These protections may sometimes impose 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 understanding how data are regulated, communities may be more effective in accessing data needed to promote community health. For example, a community that understands what data are and is not 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 is designed to provide users of community health data an introduction to these systems.

Appendix B: Federal and State Laws

Human Subjects Research

A brief summary of the regulation at 45 CFR 46, Protection of Human Subjects, also known as the "Common Rule," is provided to prompt community groups using health data to consider whether projects must comply with this federal regulation. The most authoritative primary source of information about federal human subjects regulation is found at on the web site 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 conducting 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. 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 is an example of manipulating a person's environment.

Interactions include any communication or contact between a data collector and the person, which occur, 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 that a person provides for specific purposes and that are expected to remain private (for example, a medical record). Information a person provides to a reporter would not be private information. If the information is not linked to a specific person who is identified 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 the purpose of 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 establish 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 typically *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

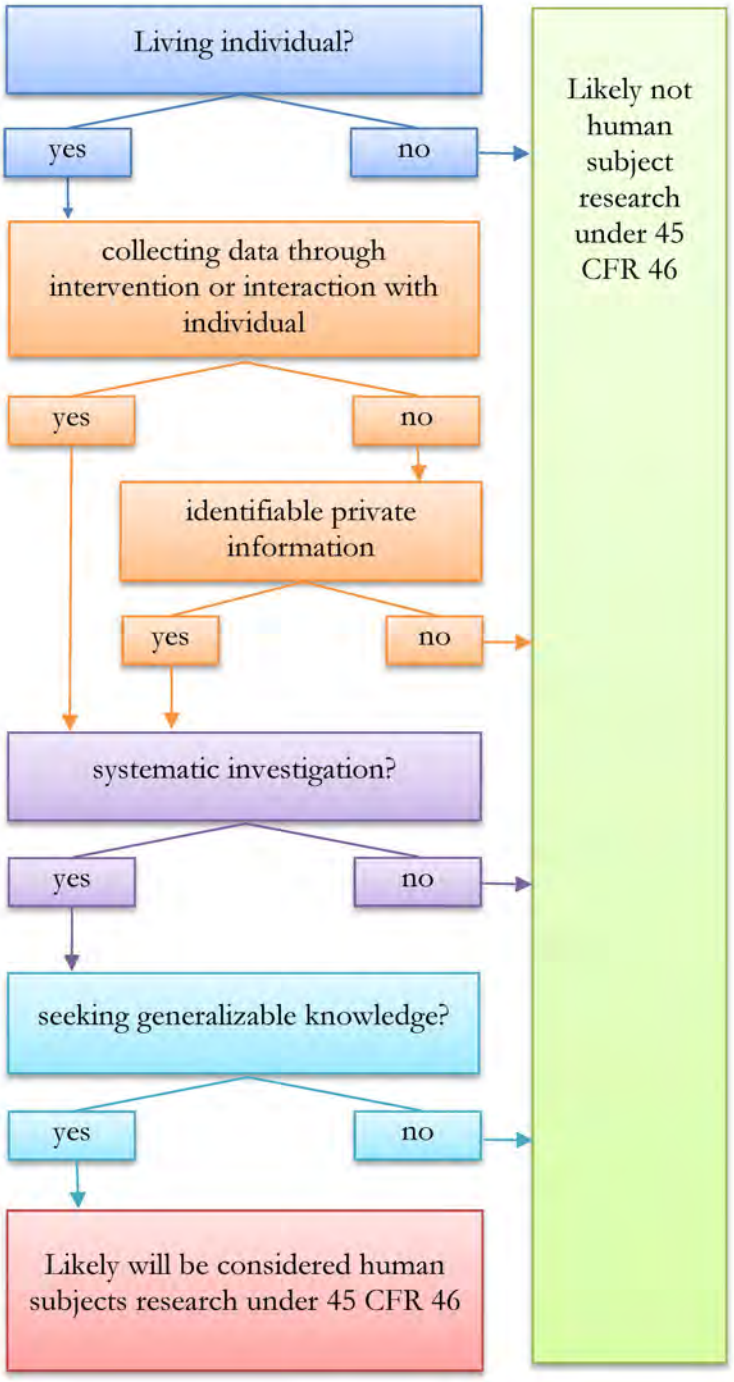
It may be necessary for an IRB to review a proposed project to assure that these activities are not research under 45 CFR 46.

Next Steps

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 laws and regulations governing their project.

Appendix B: Federal and State Laws

Is a project human subject research?

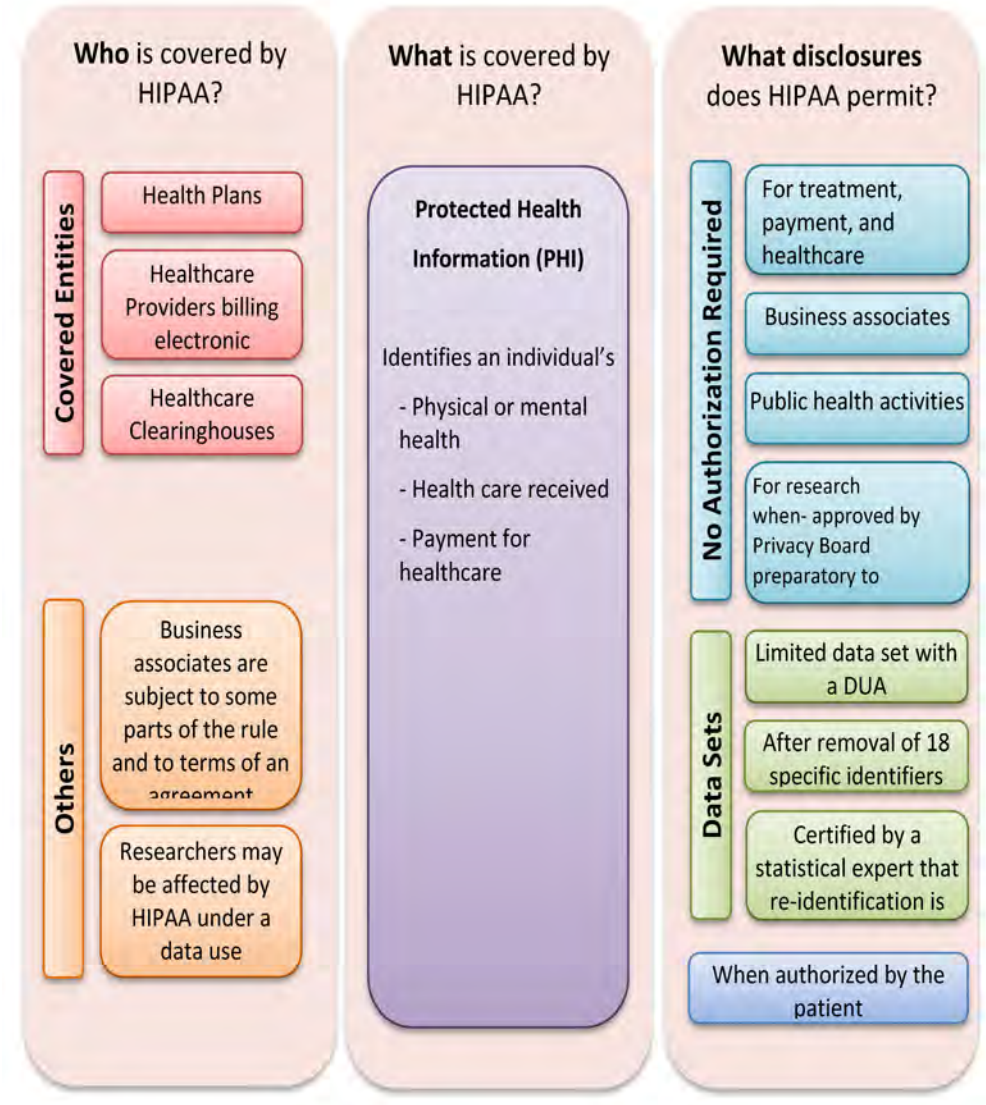


Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule

A brief summary of the HIPAA Privacy Rule is provided to prompt community groups using health data to consider whether they may be covered or to understand the obligations of entities providing data to them.[11]

Health data users should know:

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



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

Appendix B: Federal and State Laws

An entity seeking to use data from a HIPAA covered entity (broadly speaking, health care providers, insurers, and health care clearing houses) may need more information than is provided 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 permitted 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 to be 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	Device identifiers and serial numbers
*Geographic subdivisions smaller than a state	Web universal resource locators (URLs)
*Dates	Internet protocol (IP) address numbers
Telephone numbers	Biometric identifiers, including fingerprints and voiceprints
Fax numbers	Full-face photographic images and any comparable images
Email addresses	Any other unique identifying number, characteristic, or code, unless otherwise permitted by the Privacy Rule for re-identification
Social security numbers	
Medical record numbers	
Health plan beneficiary numbers	
Account numbers	
Certificate/license numbers	
Vehicle identifiers and serial numbers, including license plate numbers	

*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 advance 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, such as the one found at Appendix E, used by CMS when releasing 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 mechanisms 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 mechanisms for engaging the community, they have different approaches to protect individual privacy and confidentiality. The eMERGE network is engaged in a continuing effort to define what it means to de-identify biospecimens, biological data, and clinical information.

Vanderbilt

Vanderbilt's system involved a web survey of 4037 individuals and a community advisory board. The Community Advisory Board was established to ensure that the community had a voice. Board members interacted 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 background. Vanderbilt found community board members to be inquisitive and active participants. They were not passive; rather they wanted to know about what the eMERGE network was doing, and they wanted to provide recommendations.

Vanderbilt also found that community boards alone were not enough: community members 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 interacting with the community and assure that concerns reached 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 imposed additional layers of oversight, including evaluation by the University's Ethics Committee and three oversight boards: Ethics, Scientific, and Community Advisory. Their de-identified repository allows individuals to opt out of participation. In addition researchers using eMERGE data must register each study separately alert researchers when their data use may violate policies and the wishes of the members of the community whose data are being used.

Sources:

Bradley Malin, PhD, Vanderbilt University (testimony and correspondence).

eMERGE web site, http://emerge.mc.vanderbilt.edu/.

Mayo Clinic

When the Mayo Clinic started biobanking and reuse of the electronic medical record system, it adopted a deliberative democracy model. The model engaged community members in open dialogue for four 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.

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 Network web site, 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: 1001-1007.

Newborn Blood Spots

Almost every baby born in the U.S. is screened for a range of diseases by taking 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 with the potential to 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. While 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 had the potential to be 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 have the option of opting 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 web site where parents can learn more and complete the process of opting in or out of research. This web site serves as a good example of how data users can promote openness, transparency, and choice.

Sources:

Botkin, JR, Goldenberg, AJ, Rothwell, E, Anderson, RA, Lewis, MH. **Retention and Research Use of Residual Newborn Screening Bloodspots.** *Pediatrics.* Jan 2013; 131(1): 120–127. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3529945/

Michigan Department of Community Health, Consent Options, 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 having a 15-minute block of 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 forget about how management gurus say a meeting should be run; rather "just follow your grandmother's advice: sometimes you just need to listen and not say anything."

Source: Testimony of Dr. Phillip Smith, IHS IRB, NCVHS Subcommittee on Privacy, Confidentiality and Security, April 17, 2012.

A Refugee Community's Expectations

One community health promotion project found that members of 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; rather, they wanted the family to complete the survey as a unit. Although the organization's IRB found this approach disturbing because it would not preserve confidentiality among family members, they agreed to proceed in accordance with community members' preferences.

Source: Linda Silka, PhD, 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 investigate 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 community members in creating the survey, community members informed the development of surveys, and community members 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.

Law 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, PhD, University of Colorado Denver (interview and correspondence)

PINE Study

The PINE Study is the product of collaboration among the Chinese Health, Aging, and Policy Program at Rush University, Northwestern University, and over twenty community services organizations, including the Chinese American Service League and Xilin Asian Community Center as the main community partners. This academic-community partnership is guided by community-based participatory research (CBPR) 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 harsh violence and decimation facing Chinese community in the past. The issue is further compounded by vast cultural and linguistic barriers.

Between 2011 and 2013, the PINE Study conducted face-to-face interviews with 3,159 community-dwelling older adults between 60 to 105 years old. The multilingual staff interviewed participants according to their preferred language 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 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 the average older adult in the U.S. Many experience low acculturation levels, financial hardship, and insufficient social support. 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.

Source: Dong X, 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 Jan 3;2010:124246.

MyHealth Access

"MyHealth Access Network is a non-profit coalition of more than 200 organizations in northeastern Oklahoma, with a goal to improve health care quality and the health of area residents while controlling costs. Our organization was chartered to facilitate communications and connections among participants in the health care systems. As such, MyHealth does not directly provide care, but provides those who do with technology, information, communications, and analytics to support improved care quality and reduced costs." http://www.myhealthaccessnetwork.net/

MyHealth Access Network engaged the community in a 100-day planning process that involved 200-300 people. At the outset, 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 finding 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 explains the widely recognized success of the model.

Source:

Interview with David Kendrick, MD, MPH, MyHealth Access

Research on biological samples from members of the Havasupai Tribe

Members of the Havasupai tribe provided DNA samples to Arizona State University 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 members of the tribe. This illustrates that harm is not only caused when personal health data are disclosed (as in the hospital discharge data set), but when every member of a small group can be stigmatized.

After a law suit was filed, ASU agreed to a settlement to "right the wrong" in using the data in a way that violated the rights of tribal members whose right to consent was violated.

Source: American Indian and Alaska Native Genetics Resource Center web site, 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

Appendix D: Worksheet and Check Lists

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

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

Describe plans to mitigate 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

Appendix D: Worksheet and Check Lists

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 in accordance with acceptable data collection and use practices?

- ☐ If the organization lacks expertise in data collection best practices, seek 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 established and documented in advance of 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.

Appendix D: Worksheet and Check Lists

Data Analysis

Accountable individual/entity: _____

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

Reporting Results

Accountable individual/entity: _____

Describe how reported results will protect communities, subgroups, or individuals from bias or stigma, and describe protections to assure accurate reporting of results.

Describe protections to assure accurate reporting of results.

Appendix D: Worksheet and Check Lists

Data Security

Accountable individual/entity: _____

Identify mechanisms for protecting 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

Appendix D: Worksheet and Check Lists

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)

Appendix D: Worksheet and Check Lists

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 permitted in the scope of the agreement
- ☐ Require recipient to report to the data source any use or disclosure not permitted 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, understand:

- ☐ 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 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 obligations 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?

Appendix D: Worksheet and Check Lists

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:[12]

- ☐ 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 DUA.
- ☐ LDS recipient agrees to report LDS use or disclosure 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 use appropriate safeguards to prevent use or disclosure not specified in the DUA.
- ☐ LDS recipient agrees to refrain from identifying or contacting individuals whose health information is contained in the LDS.

[12] The University of Wisconsin has compiled a HIPAA Privacy Rule Research Guide that may be helpful, including a checklist similar to the one above. The Research Guide may be found at https://hipaa.wisc.edu/ResearchGuide/.

Appendix E: Data Use Agreements

[placeholder for pointers to example agreements]