Let's Talk Data!
Getting Comfortable with Data Lingo

When talking about data it is important to have a shared vocabulary. This ensures that everyone involved in collecting, analyzing, reporting or sharing data has a common understanding about the process. This fact sheet includes common words and terminology used to describe types and levels of data (qualitative and quantitative) as well as measuring, accessing and sharing data.

But first, let's look at an example of a county interested in using their data to assess progress:

In the last five years, L'Enfant County[1] has noticed that babies born in their county have more health issues than before. County officials have met with local hospital leadership and pediatricians to see how best to address this issue. They decide to launch a new initiative called Healthy Parents, Healthy Babies. The county is hoping this initiative will increase timely prenatal care and improve health outcomes of infants and toddlers. To track their progress on this outcome, L'Enfant County must use data from other state and local agencies.

TYPES OF DATA

Raw Data

- The "original data." Not yet manipulated, transformed, or otherwise altered. Raw data can include both data collected using paper and pencil, as well as data that has been transferred to an electronic database.
- Raw data is collected from the primary source. This could include data collected by a program or organization, or data collected for a research study.
- Raw data has not been cleaned yet for use. It may need little to no cleaning or may need extensive cleaning.
- Raw data may contain personally identifiable information or already be de-identified (see descriptions of personally identifiable and de-identified data below).

For example, if L'Enfant County wanted to track individual attendance at a parent class, they would request the parental attendance records, which may come in the form of the sign-up sheets for the classes; this would be an example of raw data.

Administrative Data

- Typically refers to information programs collect about children, families, or staff delivering programs. Data are usually collected to meet legal requirements, or to meet the requirement of a particular program or funder.
- The primary use of administrative data is for record keeping. Other uses include determining eligibility, tracking enrollment, monitoring staff workload, and documenting services provided/received.
- While these data are usually collected for tracking and reporting requirements, they can be useful beyond that purpose. They can be used to help answer important program, policy and research questions.
- Administrative data can also include health and behavioral data, such as results of screening tools, diagnoses, and treatment plans.

1 L'Enfant County is a fictious county. All information and data presented in this document about L'Enfant County are not real.

- Programs often keep these data in multiple places or databases.
- Sometimes administrative data are collected and stored at the individual level of a child or family, meaning that each child has an identifier and an entry in a database. Sometimes it is stored at the aggregate level, such as numbers of people served, reached, or trained.

In L'Enfant County, their new initiative, Healthy Parents, Healthy Babies, is providing expectant parents with a series of free classes on infant care. Total enrollment for each class (aggregate administrative data) is being tracked to assess the demand for these classes. Additionally, individual attendance is being tracked to determine whether parents are receiving the full curriculum, which is an example of individual administrative data.

Qualitative data

- Non-numerical, descriptive data that can be grouped into categories.
- Qualitative data is often collected through observations, interviews, and focus groups.

Policymakers in L'Enfant County conducted a focus group to better understand families' difficulties in finding health care. In reporting the results, they grouped responses into three different categories: transportation issues, lack of specialists, and financial difficulties.

Quantitative data

- Numerical data, such as a measurement, count or score.

Policymakers also collected quantitative data on time spent finding a specialist, distance from county residents' homes to the nearest prenatal specialist, and resident's numerical rating of satisfaction with their health care options.

LEVELS OF DATA

Population-level

- Data were collected on everyone in the population.
- OR data are representative of a given population. That is, data were collected on a sub-set of the total population but represents the total population.
- Examples include a census or nationwide survey, or a survey, interview or measure used to collect data on a given population.

In L'Enfant county, a survey was administered to every household with children under age three (i.e., at the population-level) to assess whether families felt they had adequate access to prenatal care. Since the county is small, administrators were able to reach a 100% response rate from every family with a child under three. However, if the response rate was lower, for example a 25% response rate, it is possible the data could still be considered at the population-level if it accurately represented the demographics of the total population. In order to determine if their data were representative of the population, the county could work with a research partner to determine population-level representative estimates of their population and compare those to the response rate.

Individual-level

- Data are collected and stored on individual children, teachers and caregivers, or families.
- Data are not necessarily identifiable. Yet, there is sometimes enough demographic information available for each child or family in individual-level data that the data could become identifiable. For instance, even when a dataset does not contain a name or address, if there is enough information attached such as date of birth, race/ethnicity, gender, or neighborhood, then it might be possible to identify that individual. Whenever using individual-level data, it is important to follow the correct data security and confidentiality protocols for those data.

In L'Enfant County, a birth preparation class records the stage of pregnancy at the first visit for each participant. Data for each participant is then stored in a database where there is a separate record for each person.

Aggregate-level

- Data were collected at the individual level but have been combined across individuals.
- Aggregate data can be combined into a population or sub-population level. For instance, data can be aggregated up to the county level (e.g., total number of infants and toddlers in the city), or up to different sub-populations (e.g., total number of infants and toddlers who were born to teenage mothers).
- These combined numbers are considered aggregate if they are large enough to make identifying specific children or families impossible.

L'Enfant County maintains a database with information on enrollment in the Supplemental Nutrition Assistance Program (SNAP). This database has individual-level data for each family enrolled. A researcher was interested in knowing the total number of families by school district who were enrolled. So, the county provided the researcher with a separate spreadsheet of these totals aggregated by school district, so that individual families could not be identified from the data.

Disaggregated

- Data that is broken out by subgroup characteristics.
- This data can still be aggregate level, e.g. percentage of babies born at low birth weight by race and ethnicity.

The researcher wants to know whether SNAP enrollment differences by school district might be explained by the accessibility of enrollment materials, which are only available in English. She requests disaggregated data on SNAP enrollment by primary language.

Longitudinal data

- Data that are collected on the same (or very similar) variables, from the same sample, with repeated measurement over a time period.

Babies in L'Enfant county are weighed and measured every three months from birth for a period of two years as part of a project to track growth and development.

MEASURING & ACCESSING DATA

Data source

- The entity that collects and manages the data. This could be an agency, organization, county, city, state, etc. Sometimes this is the same entity that funds the collection of data, but sometimes it is multiple entities.
- Crediting the data source is important when reporting findings.
- When interested in using or requesting data, reach out to the entities that funded and collected the data, which are sometimes the same entity.

The Office of Child and Family Health for L'Enfant County collects records of preterm births in hospitals throughout the county, and therefore would be the data source for these data.

Measure

- The instrument used to collect data. This could be a survey or tool (e.g., Ages and Stages Questionnaire), interview protocol, etc.

A nurse at a clinic in L'Enfant County wants to assess parent confidence in their ability to care for their newborn. He develops a series of questions asking each parent to rate their perceived ability to manage various tasks, including feeding, diapering, and managing sleep schedules. This series of questions is administered as a set, known as a measure, called the Parental Confidence in Newborn Care Questionnaire.

Data element

- Smallest named item of data that conveys a specific meaning. Sometimes called a 'variable', it is the attribute of the data unit.

Data elements in the L'Enfant County Infant Health database include variables such as birth weight, home zip code, and number of prenatal health clinic visits.

Data collection

- The planned use of measures or observations to gather information.
- Data collection can occur in one or many settings. It can also occur at one time period, in conjunction with specific events, or at regular intervals.

As part of an initiative to increase access to prenatal care, L'Enfant County is conducting a survey of transportation options available to expectant parents. Families receive a survey in the mail asking them about the feasibility of using public transit to access health clinics. Findings from the survey are one type of data collection that the county is using to understand access to prenatal care.

Codebook or Data dictionary

- A document that lists variable names, what they mean, how they were measured, and any other information that would help someone understand a variable.

An analyst wants to test for differences in infant health measures by geographical location. The database includes several potential variables titled "location," "zip" and "region." She guesses that "location" will tell her where a participant lives, but after consulting the data dictionary, finds that location refers to the place where measurements were taken. The definition for "region" is better suited to her needs.

Data access

- The ability to view, process or manipulate a set of data or data collection elements.

L'Enfant County healthcare administrators want to know the average number of complications per birth in the last five years. They request access to a set of aggregate data from local clinics that have been collecting this information in order to track birth complication rates over time.

Primary data

- Primary data are data collected by an organization for a specific purpose. These data are generated through an organization's planned use of measures or observations.

The county believes that cultural competence of area clinicians might be a barrier to accessing timely prenatal health care for a population of recent immigrants to L'Enfant County. The county contracts with a local researcher to conduct a survey of clinicians' familiarity with this population's cultural practices around birth as well

Data can be collected and used for multiple purposes. It may be used for programmatic and reporting purposes such as monitoring receipt of child care subsidies. It can also be used to track use of services and referrals such as in a case management system. Finally, it can be collected to answer a specific policy or research question. The reason for collecting data often determines how it is gathered, managed, and analyzed. However, sometimes data collected for one purpose can be used for another, such as linking administrative data to answer a policy or research question.

as language and translation options. The data are being collected for the first time in an original context and are primary data.

Secondary data

- Data that have already been collected for a specific purpose by a certain entity that is now being analyzed by a different entity.
- The purpose of the secondary data analysis is usually different from the purpose for which the primary data was collected; however, any analysis of data not collected by the organization analyzing the data is secondary data analysis.
- Secondary data analysis can also include re-analysis of data, but only if that re-analysis is for a different purpose than the original primary data analysis.

L'Enfant county administrators know that a local library developed a program to provide health information to pregnant teenagers. To evaluate this program, they recorded which materials expectant parents used from a special collection geared toward prenatal and family health. The library's goal was to track how often the collection was accessed compared to health information that was not specially curated. L'Enfant county administrators now want to use this data to understand whether internet, pamphlet, or book-based media are preferred by teenagers seeking health information about pregnancy.

Common identifiers

- Markers or variables that help to identify cases or data stored in separate locations. This can be data from different databases, or multiple entries within a database.
- Can be an identifiable variable like a name or birth date, or a keyed variable like an identification number.

When an expectant parent attends a birth planning class, they are given a unique number that they are to enter every time they complete a survey on their mobile phone. The study staff keeps a list of names matched to ID numbers so that when classes are over, they can match attendance across different classes.

Data linkages

- The process of relating data across databases using a common identifier.

Administrators want to know about program utilization across birth planning, prenatal screening, and health education interventions. Because these programs are all provided by the same agency, they can pull data from separate databases using a common identifier to link an individual's records in multiple datasets.

Public records

- Records collected by a government agency that are legally required to be available to the public. This can include birth records, census data, and records of government meetings.
- "Available to the public" does not necessarily mean "easily available." Finding public records may require visiting government offices, paying for reproductions of documents, or submitting detailed requests.
- For more information on using public records in research see:
  - https://ctb.ku.edu/en/table-of-contents/assessment/assessing-community-needs-and-resources/public-records-archival-data/main
  - https://www.archives.gov/research

In L'Enfant County, residents with drug convictions can be denied vouchers for Section 8 Housing. County administrators submit a request for criminal records of these convictions for parents involved in a study of housing instability to assess the scope of this problem.

DATA SHARING AND PRIVACY

Data sharing

- Providing access to data.
- Making data useful to another group.

The L'Enfant County administrators securely sent data to a researcher for use in a study of access to prenatal health care.

Data sharing/use agreements

- Legal documents that are created by entities that allow them to share or use each other's data for specific purposes.
- Agreements spell out why data are being shared, common terms, partner roles, how data will be uploaded, stored, and accessed, who can access the data, and how data can be reported out.
- Agreements commonly include sections on misuse of data, data security violations, and how to amend the data sharing agreement.
- Often these documents are called data sharing agreements, data use agreements, or memoranda of understanding.

Before sharing their data, county administrators wrote up an agreement with the researcher. The agreement included information such as which data points she'd have access to, how the data will be used and how it should be stored, what kinds of reports they would want from her analysis, and the roles of all partners.

Data security

- The technology and systems that protect the data.
- Can include password protection, locked locations, and training for anyone using the data.
- Includes both storage and transmission of data.
- Data security policies are requirements and practices that an organization uses to protect data. The organization wishing to receive the data may need to illustrate their organization's ability to adhere to the policies of another organization. For example, an organization may need to agree to train their staff or provide specialized servers.

Survey responses are kept in a locked cabinet before being entered into a database. The database itself is password protected and only visible to select study staff who have completed a data management course.

Data storage

- The location where the data is kept. This can include a locked cabinet where paper documents are stored, a drive on a network, or space on a server.

Once data is de-identified, study staff upload the file to a Dropbox storage site in a password protected location. Paper copies of survey responses are kept in a locked cabinet, and databases with identifiable data are stored on a secure server.

Institutional review board

- An administrative group at an organization that works to ensure research is conducted ethically. The review research plans and ensure that rules about privacy and confidentiality are followed.
- Typically research organizations, and sometimes state and local agencies, have an internal institutional review board (IRB), or are attached to one that they use.

In order to assure that administrators were conducting their research ethically, they utilized the services of an independent institutional review board to review their work.

Privacy

- Data privacy is one's right to know what data are being collected, for what purpose, and who has access to that data.
- This can vary by setting. For instance, states, communities, programs and agencies may all have their own requirements around data privacy and what types of data can or cannot be shared with an external person or entity.

- Depending on the data and where they are stored, there can be federal regulations the data fall under regarding privacy restrictions. For example, HIPAA[2] laws discuss privacy related to health information. FERPA[3] laws cover private information and educational records in schools.
- Laws should not be the only consideration when attending to privacy around data. The information an individual considers personal can vary culturally or generationally as well.

When entering the hospital to give birth, a mother is given a "consent for services" form that details how her health information will be shared during treatment. This form is required by both the hospital as well as by HIPAA in order to protect the privacy of her data or health information. She agrees to the process detailed by the hospital but opts out of sharing her information with the hospital's research database. She feels that information about her health and the health of her baby shouldn't be shared unless medically necessary.

Consent

- Fully informed agreement by a person who possesses the capacity to understand the information relevant to a decision.
- It is important to ask about whether parental consent (if the child is too young to provide their own consent) is required to share data when entering into a data sharing agreement.
- Even if parent consent was originally obtained when data was first collected, additional consent may be needed to share data with another organization.
- For more information on the consent process, protections for human subjects when using data, and developing consent forms to access data, please go to: https://www.hhs.gov/ohrp/regulations-and-policy/guidance/faq/informed-consent/index.html

A local United Way in L'Enfant County is interested in obtaining data from a home visiting program whose goal is to connect participants to prenatal care. The United Way wanted to determine which of their families had been part of this home visiting program and in turn had been referred to services and received timely prenatal care. In order to share the data, the home visiting program required a signed consent form from parents agreeing to share their data with the United Way.

Confidentiality

- Protecting the data (identity and/or sensitive information) of individuals.
- Placing limitations on who can access and work with the data or masking the identity of an individual in a dataset if required.

Before the home visiting program was able to share the data with United Way, they set up a data sharing agreement that specified who would be allowed to have access to the data, and to which data they would have access. This ensured that there would be strict confidentiality rules in place so that only authorized people could have access to the data.

Personally Identifiable Information (PII)

- Information that can be used to identify a person.
- This may include directly identifiable information such as:
  - Name, address, social security number
  - Age or date of birth
  - Bank, credit card, or other personal account numbers
- This may also include linked information. For example, if a database doesn't have any direct identifiers but includes information that can be grouped to figure out a person's identity, that information is considered identifiable.

2 HIPAA laws regulate the disclosure of health-related information. See https://www.hhs.gov/hipaa

3 FERPA stands for Federal Educational Rights and Privacy Act. These laws protect the privacy of educational records. See https://www2.ed.gov/policy/gen/guid/fpco/ferpa

- For example, if a database identifies a classroom, a child's age, and their home zip code, one may be able to narrow down this information to a particular child.
- Unusual circumstances warrant special attention. For example, if a database identifies a person as having recently visited an emergency room and being employed with an organization, that circumstance may be rare enough to identify the individual.

A recent report of prenatal care access discussed complications in a preterm birth that occurred in a very rural area of the county. While the report did not provide any names, addresses, or birth dates, the population was so small, and the complications were so unique that readers were able to identify the child and family in the report.

De-identified data

- Identifiers of both an individual and those in their immediate environment (relatives, employers and household) are removed from a data set.
- This includes personally identifiable information (see above) such as names, certain geographic identifiers (particularly in combination with other information), telephone numbers, social security numbers, and email addresses. What is considered de-identified may depend on the specific regulations or policies. For instance, for regulations for data that fall under HIPAA, please visit: https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html#rationale

Names and addresses are removed from a statewide dataset of healthy births before being shared with L'Enfant County administrators. Birth dates are rounded to each quarter so that administrators can examine birth rates on a larger scale without identifiable information.