The way public health sends and receives data, including surveillance data, has changed dramatically over the years. Historically, providers sent paper reports to public health agencies. Later, they logged into a surveillance system to report cases. In contrast, harvesting data from an EHR system for surveillance purposes, as with electronic lab reporting, makes public health a secondary user of the data collected for health care or administrative purposes. As such, providers do not necessarily know that the data they enter in an EHR goes to public health agencies or the purposes for which public health will use the data.

Due to this lack of awareness on the part of providers, understanding the *provenance* of the data you will be receiving—in other words, who is collecting the data, in what format, for what purpose or purposes, and with what attention to quality—becomes more important than ever. With this knowledge, you can determine how much confidence to have in the data, if its quality suffices for your purposes, and whether you and the provider interpret and assign the same meaning to the individual data elements.

You earlier defined the data elements of interest to your program; you now need to better understand details about those data from the perspective of their native health care and EHR environments. Answer the following questions to gain that necessary understanding:

- Who collects the data?
    - Administrative or clinical staff?
    - How consistently across staff?

- At what point in the clinical workflow are the data collected?
    - Admission/check-in?
    - When roomed?
    - During physician-patient encounter?
    - Hours or days after patient appointment?

- Who enters the data in the EHR?
    - Administrative or clinical staff?
    - How consistently across staff?

- How are the data obtained?
    - Using an instrument or device?
    - Via a verbal report from patient?
    - As an estimate from a staff member (e.g., "seems about 6 feet tall")?
    - From a lab result?

- Are the data structured or unstructured?

- o If structured, using a national vocabulary standard or a proprietary EHR coding scheme (see below and the *Defining Data Elements* tool in this section)?
  - o Is staff sufficiently trained to pick the appropriate code?

- For what purpose(s) are the data collected?
  - o To meet Meaningful Use requirements?
  - o For billing?
  - o For quality reporting?
  - o For diagnoses?
  - o Does the person collecting or entering the data, or performing both tasks, know why they are being collected?

You could ask the six questions above for every data element of interest to you (see the *Understanding Clinical Data and Workflow* worksheet). Not all data elements will be equally important for your project, but learning more about clinical workflows and the business of health care can only help you to understand the strengths and limitations of the data upon which you hope to act.

Understanding structured vs. non-structured data

The way data are recorded significantly impacts the usability of the data for secondary purposes. Clinical data stored in EHR systems are typically either structured or non-structured:

- Structured data follows a prescribed data model and value set, constraining the users to only be able to entering or choose pre-determined values. Computers can readily process structured data, which it often stores in databases. Data sent by medical devices to EHRs will typically send structured data.
  Example structured data entry: SNOMED CT (Systematized Nomenclature of Medicine—Clinical Terms) is one of many structured code sets used to indicate medical phenomena. Within an EHR, the recorder selects, "Type I diabetes mellitus without complication (disorder)" from a list. The system stores the SNOMED CT code 313435000, a value that has the same meaning across the United States (see *Semantic Interoperability* in the *Analyzing Technical Options* section for more information on code sets and vocabularies used for structured data).
- Unstructured data (also called "free text") does not follow a pre-defined set of values, allowing users to instead enter narrative information about data using their own words. This means of recording data provides the user with the most freedom for recording an entry, but because the same clinical event could be documented in myriad ways, computers can't easily process unstructured data, making errors more likely.

Example free text entry: The recorder enters, "The patient has diabetes with no complications." Note that a recorder could enter this in many different ways. Secondary use would require parsing the text and trying to fit it into a category. While this task may be trivial for a human reader, it can be very challenging for a computer to perform. As a result of this issue, much of the *natural language processing* research in health care now focuses on accurately interpreting free text entry.

Structured data is key to consistency in data analysis and to support computer analysis. In addition, careful definition of the concepts and the values that represent the data elements can ensure compatibility of data from different systems, or at a minimum, provide a way to map them to each other. In a collaborative project with multiple community partners, taking this approach assumes particular importance.