Appendix 3B: Considerations for Data Elements

Any member of the birth defects surveillance program staff involved in selecting, maintaining, or utilizing data elements should be familiar with how data characteristics impact program activities. They should also understand the origins of the data elements, the various formats for data elements within the birth defects surveillance database, and strategies for finding the location of data in the medical record to populate each data field. These aspects of data criteria and data elements are described below.

3B.1 Data Element Characteristics

Data element characteristics, as detailed below, influence the amount of effort a birth defects surveillance program will expend on data collection and quality assurance measures, while also impacting the statistical analysis process and other program outputs.

- **Availability:** Data must be available to the birth defects surveillance program and retrievable from the data source(s). Many data elements are collected and stored at the data source in clinical and administrative databases, facilitating availability and retrievability. In most cases, information should only be collected if it is consistently available. This is particularly true if the information is to be used for statistical analyses or for identifying or contacting case families (e.g., for referral to services). At times, a piece of information may be only occasionally found in the data source (and thus time-consuming to seek out and collect), but the added effort is warranted if the information is important to the case ascertainment process or other surveillance program operations (e.g., recording that the infant is in foster care or has been placed for adoption).
- **Consistency:** The format of collected information should be fairly consistent from source to source, and case to case. This is especially important for data that the birth defects surveillance program accepts as-reported from the data source, and for data mining. Simple issues such as the content of a data field, or even the field size, can significantly affect the consistency of the data and thus impact the usefulness of the data. The rules and practices for birth defects diagnosis coding are a special area of concern (see **Chapter 4** and **Chapter 5**).
- **Accuracy:** The information collected should be accurate. If the information is of questionable veracity, then it should not be collected. Utilize reliable data sources, use logic checks, and compare between data sources to ensure their information matches.
- **Uniqueness:** Birth defects surveillance programs should avoid the collection of redundant information. Information should not have to be recorded in more than one field. For example, if the pregnancy outcome date and the mother's date of birth are collected, then the mother's age at delivery does not need to be collected since it can be calculated from the existing information instead.

- **Definability:** There should be a clear definition for each data element a birth defects surveillance program collects, to ensure that there is no ambiguity about the type and format of information that a data field requires (see **Appendix 3A**).
- **Collectability:** The data elements should lend themselves to easy abstraction from the medical record. Complex or subjective information takes excessive time to locate and collect, may be of dubious quality due to inter-staff variability, and may require extensive efforts for quality control.
- **Comparability:** The birth defects surveillance program may want to consider whether other birth defects surveillance programs have access to analogous data sources and data elements. Using a unique data source or collecting a unique data element may prevent a birth defects surveillance program from comparing its data to those of other programs. Thus, the use of standardized sources may help birth defects surveillance programs pave the way for pooling data at a national level.

3B.2 Data Element Origins

A data element's origin refers to where it came from, meaning how the birth defects surveillance program obtained the information to enter in that data field. Data elements may be reported, abstracted, derived, or created.

- **Reported data elements:** These are data that are directly reported to the birth defects surveillance program by the data source (e.g., data elements available through the vital record).
- **Abstracted data elements:** These are data that are only available through abstraction of medical records received directly from the data source.
- **Derived data elements:** These data elements are not collected directly, but are rather derived by the birth defects surveillance program based upon information collected from the data source (e.g., census tract numbers, standardized geographic tables, diagnosis codes assigned by the birth defects surveillance program upon review of the medical records).
- **Created data elements:** These data elements are created by the birth defects surveillance program for the sake of internal record-keeping and/or workflow (e.g., unique case identification numbers, staff ID numbers).

Some data elements may fall into more than one of the above categories. For example, if the mother's age at delivery is not available from the data sources, it may be derived using the date of delivery and the mother's date of birth. The origins of each of the Core and Enhanced data elements are summarized in the *Data Source* row of each table in **Appendix 3A**.

3B.3 Data Element Formats

Common data field formats are described below. The format for each of the Core and Enhanced data elements is summarized in the *Data Element Format* row of each table in **Appendix 3A**.

- **Numeric field:** A field that includes only numbers.
- **Alphanumeric field:** A field with both letters and numbers.
- **Date field:** A field that includes only dates, which are comprised of month, day, and year.
- **Text field:** A field that can contain letters, numbers, and punctuation. Text fields are often of fixed width. Text fields of infinite width are often called Amemo@fields.
- **Checkbox:** A field that contains only two options: Yes (checkbox selected), or No (checkbox empty).
- **Coded data field:** Data may be collected and stored as they appear in the data source, or they may be coded (see **Chapter 4**). A code may contain numbers, letters, or both. Whether a birth defects surveillance program collects and stores coded data or not depends on the types and potential uses of the data. From the perspective of data elements, there are several potential benefits to using codes:
    - Decreases variability in the format compared to free text descriptions (e.g., "cleft lip" versus "lip cleft" versus "harelip"), which enables more efficient search through the database.
    - Facilitates timely and efficient analysis of data and referral of cases.
    - Creates consistency amongst researchers, which enables comparability between different birth defects surveillance programs using the same or comparable coding systems. This applies not only to diagnostic codes, but also to characteristics such as maternal race and ethnicity.

3B.4 Data Element Locations

Location refers to where the data element can be found, both in terms of which data source to obtain it from, and where to look within the documents from a data source.

Obtaining data elements from a variety of data sources allows for greater thoroughness in data collection. Thoughtfully choosing which data source to search for each data element can not only help with faster location of the data, but also higher accuracy of the data obtained.

If a data element is missing or questionable in one data source, it may be available or more accurately recorded in another. NBDPN recommends using the preferred data source identified in **Appendix 3A** for each data element. While there are preferred sources, it is important for birth defects surveillance program staff to use discretion as there may be instances of input error from a data source. When discrepancies exist, the following general guidance can help with decisions about which data source should take precedence over other sources:

- Vital records take precedence as the source for **demographic** information.
- The case report and additional clinical records, as needed, take precedence as a source for **clinical and diagnostic** information.
- The source closest to the point at which the specific piece of data was originally produced is generally the most accurate for that piece of data. For example:
    - The test result data directly from the cytogenetics lab is likely the most accurate for karyotype information, rather than a clinical progress note.
    - The delivery medical record is likely the most accurate for birth weight, rather than the birth certificate.
    - The pediatric ophthalmology clinic is likely to provide the most accurate diagnosis for a congenital eye condition, rather than a summary note from a primary care provider.

Birth defects surveillance programs should periodically re-evaluate which data source provides the most accurate information for each purpose.