

Chapter 4

Data Variables

Table of Contents

4.1 Introduction	4-1
4.2 Criteria to be Considered in Selection of Data Variables	4-2
4.2.1 Type of Case Ascertainment	4-2
4.2.2 Program Objectives	4-2
4.2.3 Data Characteristics	4-3
4.3 The Origins of Data Variables	4-4
4.4 The Formats of Data Variables.....	4-5
4.5 Data Variable Logic Checks.....	4-6
4.6 Data Variable Location.....	4-7
4.7 Risk Factor Variables	4-8
4.8 Data Variable Tables	4-9
4.9 References	4-12

Appendices

Appendix 4.1 Descriptions of Minimum (Core) Data Variables	A4.1-1
Appendix 4.2 Descriptions of Recommended Data Variables	A4.2-1

4.1 Introduction

The potential data sources available to birth defects programs contain a wide variety of information. Each item of information a birth defects program collects requires staff time to locate, abstract, code, and evaluate, as well as computer space to store it. Thus, due to limited resources, a birth defects program must be efficient in the scope of the information it collects and the manner in which the information is collected and stored.

In this chapter we discuss a number of issues relating to the data variables that comprise a birth defects surveillance system. In Section 4.2, for example, we discuss the criteria that should be considered in selecting the variables that will be collected by a surveillance system. In Section 4.3, we present the three possible origins of surveillance data variables; that is, variables may be abstracted, derived or created. Other topics include possible formats for data variables (Section 4.4), logic checks that can be used to ensure data fall within an expected range (Section 4.5), sources for data variables (Section 4.6), and issues concerning a subset of variables related to birth defects risk factors (Section 4.7). In Section 4.8, we introduce two tables that summarize core (Table 4.1) and recommended (Table 4.2) data variables for a birth defects surveillance system. Additional detail on each of these core and recommended variables is provided in Appendices 4.1 and 4.2, respectively.

It is our hope that the information in this chapter of *The Surveillance Guidelines* will promote and guide standardization of data elements across birth defects surveillance programs. Using standard data elements is particularly important when aggregating data for regional or national analysis. Standardization allows and supports comparisons and collaborations between states.

Whether a surveillance program is based on active or passive case ascertainment, our recommendation is that vital records information or copies (including birth, death or fetal death certificates as appropriate) be obtained. This allows the collection of some data using sources from which population-based demographic information can also be obtained.

Note that we are indebted to Lynberg and Edmonds (1994) for much of the information in this chapter.

4.2 Criteria to be Considered in Selection of Data Variables

A birth defects program should consider a number of different criteria when deciding which variables to collect. These include type of case ascertainment, program objectives, and data characteristics. Each of these criteria is discussed further below. The criteria considered in compiling the lists of core and recommended variables are summarized for each variable under the heading ‘Justification’ in Appendices 4.1 and 4.2.

4.2.1 Type of Case Ascertainment

The case identification methods used by a surveillance program may place constraints on the data variables collected. The available data source(s) for program variables are determined primarily by these methods. For example, birth certificate files usually offer limited data for diagnostic confirmation of the birth defect or a precise description of the defect. An infant’s medical record, other than the newborn record, is not likely to include data on the prenatal care received by the mother (see Chapter 6 on Case Ascertainment Methods).

4.2.2 Program Objectives

A surveillance program should limit the information collected to those items needed to fulfill its stated objectives. However, it can be difficult to determine what constitutes this essential information. Often individuals, groups, or organizations that utilize surveillance information may request data on variables that are not really needed and will not be used. One guideline a surveillance program might follow is that information should not be collected if it does not serve at least one programmatic objective.

CDC defines *surveillance* as “the ongoing systematic collection, analysis, and interpretation of health data essential to the planning, implementation and evaluation of public health practice, closely integrated with the timely dissemination of these data to those who need to know” (Centers for Disease Control and Prevention, 1988). Under this definition, it is clear there are a number of functions and objectives for which a birth defects program might need to collect data:

- *Descriptive epidemiology and monitoring.* Data can be examined to determine and describe the distribution of a disease (condition) within a population along the parameters of place, person, and time. Monitoring offers quantitative estimates of the magnitude of the disease.
- *Research.* Data can be used to test hypotheses or in planning research to learn the causes of a disease.
- *Service/planning.* Increasingly, surveillance programs are using information on newly identified children with birth defects to refer them for services. These include specialized medical care, educational and early intervention programs, and genetic counseling. Data can also be applied to evaluate services and prevention measures within a population. Knowledge about the disease or condition and changes in the population can assist in optimizing available resources and services.
- *Linkage.* Variables may be used to link to other databases such that data in those databases may be associated at the case level to complement and enrich case-specific data. Linkage is also an essential surveillance management tool needed to identify and consolidate duplicates.

4.2.3 Data Characteristics

Among the important data characteristics a surveillance program should consider are availability, consistency, accuracy, uniqueness, definability, collectability, and comparability. We discuss each of these in turn below.

- *Availability.* Data must be retrievable from the data sources and be available to the birth defects program. Many data variables are collected and stored at data sources 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. If information can be found only in a small portion of the data sources, then staff will spend considerable time looking for unavailable information. The birth defects program may want to either limit collection of such information or work to identify a data source where the same information is consistently available. An exception to this may be where the information is important even if it is only occasionally found in the data sources (e.g., the fact that the infant is in foster care or has been placed for adoption). However, as noted before, this information may be difficult to find and time-consuming to collect.
- *Consistency.* It is important that the information assembled within the surveillance system has a consistent meaning from report to report. When obtaining information from a range of data sources, it is essential to have a usable level of consistency from source to source. This is especially important for passive data collection and data mining. Simple issues, such as field content and even field size, can significantly affect the comparability and usefulness of the data. Coding rules and practices are special areas of concern.
- *Accuracy.* The information collected should be accurate. If the information is of questionable veracity, then it should not be collected. Second-hand information found in medical records may be incomplete or inaccurate. If information such as medication use and exposures is important, it should be collected from a reliable source, such as through direct contact with the mother, rather than from medical records.
- *Uniqueness.* Programs should avoid the collection of redundant information. Information should not have to be recorded in more than one field. For example, if the infant or fetus delivery date and the mother's date of birth are collected, then the mother's age at delivery does not need to be collected.
- *Definability.* There should be clear definitions for each of the data variables a birth defects program collects.
- *Collectability.* The data variables should lend themselves to easy abstraction. This is a potential problem with complex or subjective information. If it takes an excessive amount of time to track down and collect the information, or if there is a high degree of inter-staff variability in how the information is collected, then the information recorded in the birth defects program's database will be of dubious quality and reliability (Horwitz and Yu, 1984; Demlo et al., 1978). In addition, extensive efforts may be necessary for quality control.
- *Comparability.* The birth defects program may want to consider whether other birth defects programs have access to the same sources and types of data. If the program uses a unique data source or collects a unique data variable that other birth defects programs do not, then the program may not be able to compare its data to those of other programs. This may be of limited importance, however, if the data are being collected to meet specific programmatic objectives, where comparison between different states or programs is unimportant.

4.3 The Origins of Data Variables

Data variables may be abstracted, derived, or created.

- *Abstracted data variables.* These are data that are available only from the data sources, and the data sources must supply them.
- *Derived data variables.* Some data variables are not collected directly from data sources but are rather derived from other information collected from the data sources, e.g., census tract numbers, standardized geographic tables, disease codes.
- *Created data variables.* Some data variables may need to be created by the birth defects program, e.g., unique case and staff IDs.

Some data variables 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 recommended variables are summarized under the heading 'Source' in Appendices 4.1 and 4.2.

4.4 The Formats of Data Variables

Data may be stored in a computer database in a variety of formats, including as a numerical field, a date field, a text field, a checkbox, or a coded data field. Each of these formats is briefly described below. The format for each of the core and recommended variables is also summarized under the heading ‘Type’ in Appendices 4.2 and 4.2.

- *Numerical field.* A field that includes only numbers.
- *Date field.* A field that includes only dates, which are comprised of month, day, and year in a variety of orders and combinations.
- *Text field.* A field that can contain letters, numbers, and punctuation. Text fields are often of a fixed width. Text fields of infinite width are often called Amemo@fields.
- *Checkbox.* A field that contains only two options – yes/no, on/off.
- *Coded data field.* Data may be collected and stored as they appear in the data source, or they may be ‘coded’. A code may contain numbers or letters or both. Whether a birth defects program collects and stores data as coded or not depends on the types of data, as well as on potential uses.

If a birth defects program plans to use a field for analysis, then it is important that the field be easily coded or categorized, permitting ready analysis rather than having to sort through a large collection of free-form text. This is because information such as race/ethnicity, diagnoses, and conditions can be described in a number of different ways. For example, a person may be described as ‘African-American’ or ‘black’. A ‘cleft lip’ may also be described as a ‘lip cleft’ or a ‘harelip’.

Coding eliminates the problem of having to sort through a variety of differing descriptions. It allows for timely and efficient analysis of data and referral of cases. Coding also enables researchers to know that they are talking about the same thing, and it allows for comparability between different birth defects programs using the same or comparable coding systems.

Whenever possible, a birth defects program should use coding systems consistent or compatible with those used by other groups, particularly other birth defects programs, thus allowing for efficient comparison of data. This applies not only to diagnostic codes but also to characteristics such as maternal race and ethnicity.

4.5 Data Variable Logic Checks

Errors may occur in the data collection by a birth defects program, either because of errors in data listed in the data source or because of errors in abstraction. A birth defects program should have some method to identify and correct errors (see Chapter 7 on Data Quality Management). One means of identifying and correcting errors is through *logic checks* that ensure data occur within expected ranges.

Many of the core variables in a birth defects surveillance system have a limited number of options or ranges of values. For example, a gestational age of 75 weeks is highly unlikely to occur. And other variables may have certain logical relationships to one another. For example, the mother's date of birth must always be earlier than the infant's date of birth.

Suggested logic checks for each of the core and recommended variables are summarized under the heading 'Checks' in Appendices 4.1 and 4.2.

4.6 Data Variable Location

A birth defects surveillance program may have access to a variety of data sources and will collect data on a number of different variables. Clearly, the same variable may be available from several sources. Abstracting data from a variety of sources allows for greater thoroughness in data collection. If a variable is missing in one data source, it may be available in another source.

Staff collecting data should know where a given data variable is likely to be found, as well as the prioritization of sources for those variables retrievable from multiple data sources, since data sources may disagree as to the value for a particular variable. For example, the infant's delivery medical record and the birth certificate might record different values for birth weight. A birth defects program should prioritize the data sources for particular variables. In the above instance, for example, a birth defects program may decide that the birth weight in the medical record takes precedence over the birth weight from a birth certificate.

For each of the core and recommended variables, the data source – as well as the location within the data source where the variable is most likely to be consistently found – are summarized under the heading 'Location' in Appendices 4.1 and 4.2.

4.7 Risk Factor Variables

Risk factors in birth defects include: conditions, illnesses, or complications during pregnancy, labor, or delivery

Selected conditions, such as maternal diabetes and thyroid disease, have been associated with increased risk for certain birth defects (Becerra et al., 1990; Khoury et al., 1989). Information on conditions and complications during pregnancy and delivery may be useful for making syndromic classifications or identifying causality of birth defects, such as diabetic embryopathy.

However, there are a large number of conditions and complications possible during pregnancy and delivery, and birth defects programs could create lists of dozens to hundreds of them. Such long lists would require additional computer storage space and training of field staff regarding where to find the information and how to collect it. Even then, confusion may ensue over which conditions and complications to abstract and subjective differences between staff in their abstraction of this information. Moreover, the information in the data sources commonly available to birth defects programs may not necessarily be consistent or accurate (Olson et al., 1997).

For all of these reasons, birth defects surveillance programs should give careful consideration to the potential thoroughness and usefulness of routine data collection regarding risk factors as relevant to their goals and objectives. In general, programs are more likely to obtain useful information on conditions and complications during pregnancy and delivery through contact with parents, as is done in case-control research studies, than through medical records abstraction.

4.8 Data Variable Tables

In the late 1980s, before creation of the National Birth Defects Prevention Network, Larry Edmonds of the Centers for Disease Control and Prevention (CDC) – along with F. John Meaney of Arizona and Susan Panny of Maryland and others – collaborated on development of a set of core data items relevant to birth defects surveillance (Edmonds et al., 1988), based on an earlier list developed by CDC’s National Center for Health Statistics. We used the list developed by Edmonds et al. as the foundation for developing the current list of data variables that the NBDPN recommends for birth defects surveillance programs, adding a number of different variables in order to reflect the fact that birth defects surveillance programs have evolved considerably since the 1980s into programs with a variety of objectives and multiple areas of interest.

The data variables in Tables 4.1 and 4.2 (as well as in their corresponding appendices) are categorized as to whether they are infant, maternal, paternal, or contact information variables. For each data variable, we also note in Tables 4.1 and 4.2 the usefulness of that data item relative to a program’s specific objectives, which may include descriptive epidemiology and monitoring, research, service and planning, and linkage capability (see Section 4.2.2. for further discussion of program objectives).

To provide a sense of the relative importance of the data variables for a new or expanding surveillance program, we have further distinguished between minimum (or core) variables (Table 4.1 and Appendix 4.1) and recommended variables (Table 4.2 and Appendix 4.2).

- **Minimum (core) variables** are those that are considered necessary to fulfill the most basic programmatic objectives and that also meet most or all of the supplemental criteria discussed earlier in this chapter.
- **Recommended variables** are those that have the potential to enhance surveillance capability or to support broader programmatic objectives.

By glancing down the column for a specific programmatic objective (e.g., ‘research’), the reader can determine – based on the relevant check marks – which elements are considered ‘core’ and which other data elements are ‘recommended’ to support a given program objective. These data variables can be abstracted using a minimum number of data sources, including maternal records, infant records, and vital records. Birth defects programs that use the passive case ascertainment approach will find the vital record particularly useful as a data source for many of the maternal core data variables.

After reviewing these lists, birth defects surveillance staff may also wish to add further data variables they consider essential for their own specific programmatic purposes.

**Table 4.1
Minimum (Core) Data Variables**

Data Variable	Descriptive Epidemiology and Monitoring	Research	Service/ Planning	Linkage
Infant				
Unique ID	✓	✓	✓	✓
Date of Pregnancy Outcome	✓	✓	✓	✓
Sex	✓	✓	✓	✓
Infant's Name				
First	✓	✓	✓	✓
Middle	✓	✓	✓	✓
Last	✓	✓	✓	✓
Suffix	✓	✓	✓	✓
Source of Report	✓	✓	✓	✓
Medical Record Number(s)	✓	✓	✓	✓
Vital Record Certificate Number				✓
Place of Pregnancy Outcome	✓	✓	✓	✓
Pregnancy Outcome	✓	✓	✓	✓
Birth Weight	✓	✓	✓	✓
Plurality	✓	✓	✓	✓
Gestational Age	✓	✓	✓	✓
Diagnosis Code	✓	✓	✓	✓
Contact Information				
Name of Responsible Party			✓	
Address of Responsible Party			✓	
Telephone Number of Responsible Party			✓	
Mother				
Mother's Date of Birth	✓	✓	✓	✓
Mother's Race	✓	✓		
Mother's Ethnicity	✓			
Mother's Name				
First	✓	✓	✓	✓
Middle	✓	✓	✓	✓
Last	✓	✓	✓	✓
Mother's Residence At Time of Pregnancy Outcome				
Street address	✓	✓		
City	✓	✓		
County	✓	✓		
State	✓	✓		
Zip Code	✓	✓		

**Table 4.2
Recommended Data Variables**

Data Variable	Descriptive Epidemiology and Monitoring	Research	Service/ Planning	Linkage
Infant				
Text Description of Birth Defect	✓	✓	✓	
Date of Death	✓	✓	✓	✓
Birth Length	✓	✓		
Apgar Score	✓	✓		
Birth Order	✓	✓	✓	
Cytogenetic Analyses Performed	✓	✓	✓	
Diagnostic Tests and Procedures Performed	✓	✓	✓	
Autopsy Performed	✓	✓	✓	
Physicians of Record		✓	✓	
Mother				
Date of Last Menstrual Period (LMP)	✓	✓		
Date of Ultrasound	✓	✓		
Gestational Age at Ultrasound	✓	✓		
Mother's Medical Record Number(s)	✓	✓		✓
Prenatal Diagnosis	✓	✓		
Mother's Social Security Number		✓		✓
Census Tract of Maternal Residence at Pregnancy Outcome	✓	✓		✓
Mother's Telephone Number		✓	✓	
Mother's Education	✓	✓		
Prior Pregnancy History	✓	✓		✓
Prenatal Care	✓	✓		
Father				
Father's Date of Birth	✓	✓		✓
Father's Name		✓	✓	
Father's Education	✓	✓		
Father's Race	✓	✓		
Father's Ethnicity	✓	✓		
Father's Social Security #		✓		✓

4.9 References

- Alexander GR, Hulseley TC, Smeriglio VL, Comfort M, Levkoff A. Factors influencing the relationship between a newborn assessment of gestational maturity and the gestational age interval. *Paediatr Perinat Epidemiol.* 1990;4:133-146.
- Becerra JE, Khoury MJ, Cordero JF, Erickson JD. Diabetes mellitus during pregnancy and the risks for specific birth defects: a population-based case-control study. *Pediatrics.* 1990;85:1-9.
- Boudjemline Y, Fermont L, Le Bidois J, Lyonnet S, Sidi D, Bonnet D. Prevalence of 22q11 deletion in fetuses with conotruncal cardiac defects: 6-year prospective study. *J Pediatrics.* 2001;138:520-524.
- Bullen PJ, Rankin JM, Robson SC. Investigation of the epidemiology and prenatal diagnosis of holoprosencephaly in the North of England. *Am J Obstet Gynecol.* 2001;184:1256-1262.
- Centers for Disease Control and Prevention. *CDC Surveillance Update, January 1988.* Atlanta, GA: Centers for Disease Control and Prevention; 1988.
- Centers for Disease Control and Prevention. Spina bifida incidence at birth – United States, 1983 – 1990. *MMWR.* 1992;41:497-500.
- Demlo LK, Campbell PM, Brown SS. Reliability of information abstracted from patients' medical records. *Med Care.* 1978;16:995-1005.
- Edmonds LD, Mackley HB, Fulcomer MC, Panny SR, Meaney FJ. A recommended set of core data items for collection by state birth defects surveillance programs. Presented at the 116th annual meeting of the American Public Health Association, Boston, Massachusetts, November 13-17, 1988. As cited in Lynberg MC, Edmonds LD. Surveillance of birth defects. In: Halperin W, Baker E, eds. *Public Health Surveillance.* New York, NY: Van Nostrand Reinhold; 1992:173-177.
- Forrester MB, Canfield MA. Evaluation of a system for linking birth defects registry records and vital records. *J Registry Management.* 2000;27:93-97.
- Hall MH. Definitions used in relation to gestational age. *Paediatr Perinat Epidemiol.* 1990;4:123-128.
- Horwitz RI, Yu EC. Assessing the reliability of epidemiologic data obtained from medical records. *J Chronic Dis.* 1984;37:825-831.
- Khoury MJ, Becerra JE, d'Alamada PJ. Maternal thyroid disease and risk of birth defects in offspring: a population-based case-control study. *Paediatr Perinat Epidemiol.* 1989;3:402-420.
- Lynberg MC, Edmonds LD. State use of birth defects surveillance. In: Wilcox LS, Marks JS, eds. *From Data to Action: CDC's Public Health Surveillance for Women, Infants and Children.* Washington, DC: Centers for Disease Control and Prevention; 1994:217-230.
- McIntosh GC, Olshan AF, Baird PA. Paternal age and the risk of birth defects in offspring. *Epidemiology.* 1995;6:282-288.

Nielsen JP, Haahr P, Haahr J. Infantile hypertrophic pyloric stenosis: decreasing incidence. *Ugeskr Laeger*. 2000;162:3453-3455.

O’Leary PO, Bower C, Murch A, Crowhurst J, Goldblatt J. The impact of antenatal screening for Down syndrome in Western Australia: 1980-1994. *Aust NZ J Obstet Gynecol*. 1996;36:385-388.

Olshan AF, Schnitzer PG, Baird PA. Paternal age and the risk of congenital heart defects. *Teratology*. 1994;50:80-84.

Olson JE, Shu XO, Ross JA, Pendergrass T, Robison LL. Medical record validation of maternally reported birth characteristics and pregnancy-related events: a report from the Children’s Cancer Group. *Am J Epidemiol*. 1997;145:58-67.

Rasmussen SA, Moore CA, Paulozzi LJ, Rhodenhiser EP. Risk for birth defects among premature infants: a population-based study. *J Pediatrics*. 2001;138:668-673.

Torfs CP, Christianson RE. Anomalies in Down syndrome individuals in a large population-based registry. *Am J Med Genet*. 1998;77:431-438.

Whiteman D, Murphy M, Hey K, O’Donnell M, Goldacre M. Reproductive factors, subfertility, and risk of neural tube defects: a case-control study based on the Oxford Record Linkage Study Register. *Am J Epidemiol*. 2000;152:823-828.