Appendix 11.1 Data Suppression

Specific birth defects are often rare events (sometimes extremely rare) leading to a set of issues that must be considered when presenting birth defects data. The public health professional must balance the potentially conflicting goals of information dissemination with protection of the privacy of persons in the community. When the number of cases in a diagnostic category within a group or stratum (such as race or sex) is small or the population from which the cases are determined is small, the risk of allowing a specific individual to be identified may be deemed too large to be acceptable. In such cases, steps must be taken to protect an individual's privacy.

The most common method of preventing the identification of specific individuals in tabular data is through cell suppression. This means not providing counts in individual cells where doing so would potentially allow identification of a specific person. Cell suppression can also be done by combining cells from different small groups to create larger groupings that reduce the risk of identifying individuals. While there are also more sophisticated data perturbation methods that use statistical noise to mask sensitive information, these are generally more suitable for use with economic or financial data than with public health data. This appendix reviews the basic methods, issues, strengths, and vulnerabilities of cell suppression. In addition to protecting privacy, prevalence information is often suppressed when concerns exist regarding possible statistical unreliability of estimates that are based on small numbers.

Suppression Criteria

The first question is whether or not to suppress. Surveillance program administrators and technical staff should be aware that standards used to suppress data may already be set in state laws or in departmental or institutional rules and regulations. It is the responsibility of surveillance staff and administrators to be aware of these standards and practice within their limits. If standards are not established, it behooves a surveillance program to establish rules that will be followed consistently. This is best accomplished with the assistance of an advisory committee, an institutional review or privacy board, or a similar body.

Suppression rules are typically based on a predetermined criterion for the number of diagnosed cases and/or the number of births in the population or subpopulation from which the cases were identified. These numbers may also be thought of as the numerator and the denominator, respectively, of a prevalence estimate. Generally, suppression rules focus on the size of either the numerator or the denominator, the ratio of the numerator to the denominator, or the difference between the numerator and denominator. However the values that trigger suppression vary greatly from one institution or place to another, and there are no set standards. In practice, the rules used vary from relatively liberal to very conservative. Suppression rules for some of the population-based data systems used to assess progress toward the Healthy People 2010 objectives are presented in Table A11.1-1.

Table A11.1-1 Data suppression rules for population-based data systems in the HP2010

Data System	Suppression Criteria
HIV/AIDS Surveillance System	< 4 cases
National Notifiable Diseases Surveillance System	Race and Hispanic origin if < 4 cases
STD Surveillance System	County: < 4 cases; State: < 6 Cases; National: None

Source: Klein et al., 2002

Each of these suppression criteria is based on simple case counts, but they vary in terms of whether the suppression is of the overall counts or by substrata such as race/ethnicity or geography. In contrast, some surveillance programs will not report data on a birth defect if the case count is less than 5, regardless of the population size, whereas many regularly report single cases. When evaluating the prevalence of birth defects or investigating potential birth defects clusters, it is often necessary to consider birth populations that may consist of small numbers of births. In this situation information on individual cases may be essential to fulfill some of the program's public health functions but should not be included in formal reports.

While reporting small numbers of cases may threaten privacy, the threat may be greatest when reporting from small populations or when the difference between the number of cases and the population count is small. This has led to suppression rules that assess the difference between the prevalence numerator and the denominator or the case count and the population size (e.g., Land, 2001). For example, given the suppression criteria requiring a minimum difference of 15 and a single case of anencephalus in a birth population of 16, the denominator minus the numerator rule would allow the data to be shown. However, in a birth population of 15 the same data would not be shown. Given the nature of anencephalus, an alternative relevant event-specific denominator may be infant deaths in the population. In that case with a birth population of 16, a single anencephalus case would not be shown unless all the infants had died. Thus, even the seemingly simple question of the relevant population to be considered may not be straightforward and should be considered carefully in deciding when to suppress.

Extent of Suppression

Having made the decision to suppress, the question becomes what and how to suppress. The solution that provides the greatest protection of privacy is to suppress an entire table whenever a single cell presents a threat, whereas the solution that provides the least protection is to suppress a single offending cell or only those cells deemed sensitive. Suppressing only sensitive cells is called *primary suppression*. However, when a single cell is suppressed, if column and row totals are provided, they can be used to compute the value of the suppressed cell. Similarly, suppressing multiple cells may allow the values of many or all of the suppressed cells to be revealed through a series of simple arithmetic solutions. This leads some agencies to practice *complementary suppression*, also referred to as *secondary suppression*, in which nonsensitive cells are suppressed in order to support the suppression of sensitive cells. If not properly done, however, the values or approximate ranges of cells in tables created with complementary suppression can also be obtained through the application of simultaneous equations (Geissing, 2001). Complex computer algorithms can be used to determine what cells must be suppressed in order to protect sensitive information. However, these algorithms are not always effective and become excessively complex in large tables (Duncan et al., 2001). One also confronts the issue of increasing data loss when large numbers of cells are used in complementary suppression.

Threat of External Data

A final issue to be considered in deciding when and how to suppress sensitive information is the potential availability of data in multiple tables. It is not enough to simply evaluate the present table with its columns and rows; one must also consider the possible availability of complementary tables. This is especially true in the era of web-based interactive information systems that generate tables for custom queries on demand. Consider a hypothetical case where, in the process of creating a table for an annual report, it was determined that cells showing pyloric stenosis counts for the black population were potentially sensitive and the decision was made to provide only the total number of cases. Subsequently it is determined that effectively suppressing the black population's case counts would require

complementary suppression of the white population's case counts. Given that the white population's data were not sensitive, they may be subsequently published in a separate table. If so, the resulting data could be combined with the original table in order to reveal the black population's data. A similar situation would arise if, to protect privacy and present all of the data, the population strata were collapsed and subsequently data for one of the strata were published.

Summary on Suppression

The more restrictive a suppression rule, the less information a given table or report will provide. The weaker a suppression rule, the greater the potential threat of revealing confidential health information. It is a question of balancing the threat to individual privacy with the public health value of presenting the data. Overall, deciding when and how to suppress birth defects information is more a social, political, and legal issue than a technical one. The technical aspects are quite straightforward, but the contextual and procedural/policy issues are likely not to be. These all need to be considered and balanced in the local context before informed decisions can be made to suppress or not to suppress data in program reports or other documents.

References on Data Suppression

Klein RJ, Proctor SE, Bouderault MA, Turczyn KM. Healthy People 2010 criteria for data suppression. In: *Healthy People 2010 Statistical Notes.* No. 24. Hyattsville, MD: National Center for Health Statistics; 2002.

Land G. Confidentiality data release rules. 2001. http://www.amstat.org/comm/cmtepc/images/Land_confidentiality%20rules.ppt. Accessed 5/21/08.

Geissing S. Non-perturbative disclosure control methods for tabular data. In: Doyle P, Lane JI, Theeuwes JM, Zayatz LM, eds. *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*. Amsterdam, Netherlands: Elsevier Science; 2001:185–213.

Duncan GT, Feinberg SE, Rammayya K, Padman R, Roehrig SF. Disclosure limitation methods and information loss for tabular data. In: Doyle P, Lane JI, Theeuwes JM, Zayatz LM, eds. *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies.* Amsterdam, Netherlands: Elsevier Science; 2001:135–166.

See also American Statistical Association: http://www.amstat.org/comm/cmtepc/index.cfm?fuseaction=1. Accessed 5/21/08.