Council of State and Territorial Epidemiologists

CDC-CSTE Intergovernmental Data Release Guidelines Working Group (DRGWG) Report:

CDC-ATSDR Data Release Guidelines and Procedures for Re-release of State-Provided Data

January 24, 2005

This report contains guidelines for implementing the *CDC/ATSDR Policy on Releasing and Sharing Data*, pertaining to the re-release of State-provided data

Table of Contents

i. Executive Summary

This report contains 16 guidelines and six procedures for implementing the *CDC/ATSDR Policy on Releasing and Sharing Data*[9] pertaining to CDC's re-release of State-provided data. Section 1 contains the background and purpose for establishment of these guidelines and procedures. Section 2 describes the characteristics of CDC data systems that are considered in-scope for this report. The guidelines and procedures in this report have been specifically prepared to address the policies and practices that CDC programs establish for re-release of State-provided data that are not already covered by a written formal data re-release procedure at the time this report is finalized.

The data re-release guidelines included in this report are described in detail in Sections 3 and 4. Section 3 includes two guidelines which pertain to the development of data agreements with State data providers. The first guideline in Section 3 requires CDC* programs to develop data agreements with State data providers, through collaboration and negotiation, in advance of receiving data from data providers. This guideline also acknowledges that in some CDC programs, States currently release data to CDC in the absence of explicit data agreements, and a process is necessary to develop these agreements even as data sharing continues. The second guideline in Section 3 lists suggested content of the CDC program data re-release plan and is based on the guidelines for protecting and releasing data that are described in Sections 4A through 4C. The nature of confidentiality protection, a suggested content element of the data re-release plan (see Guideline 2), is based on guidelines on existing confidentiality standards and procedures (see Section 4D).

The guidelines in Section 4 are grouped in three main categories: 1) guidelines representing administrative requirements for all re-releases of State-provided data; 2) guidelines that apply to re-release of State-provided data as public-use data; and 3) guidelines that apply to the re-release of State-provided data as restricted-access data. Each guideline represents a "minimum standard" for CDC programs to address when developing their program-specific release plan for re-release of State-provided data. CDC programs may wish to adopt more stringent standards than the minimum standard. Most guidelines and procedures are accompanied by a best practices statement, reflecting applicable references, resources, or selected examples of practices by CDC programs or other Federal or State programs that appear to be consistent with the guidelines. The best practices statement is meant to be descriptive rather than prescriptive. In other words, it is being left to the discretion of each Center, Institute, or Office (CIO) and their respective programs to consider for themselves whether it is appropriate to implement the listed best practice element or implement another approach. Because CDC data systems vary widely with respect to their content and format, it is important that both the guidelines and best practices allow for flexibility within the context of the principles espoused by CDC policy.

* Throughout this document, CDC should be understood to refer to both CDC and ATSDR. CDC was in the process of undergoing a major reorganization after this report was prepared. This report refers to the CDC organizational units and titles that were being used before Futures Initiative reorganization was fully implemented; thus, the names of organizational units or job titles may have changed.

Section 5 contains descriptions of practices that support re-release of data. These descriptions are meant to help facilitate CDC's adoption of the guidelines and likely will merit more discussion and consideration within CDC.

Section 6 includes the proposed deadline and steps for CDC's implementation of the guidelines and procedures and Section 7 includes recommendations for providing feedback to CSTE during CDC's implementation of the guidelines and procedures.

While this document was developed for CDC data systems having specific characteristics (see Section 2), selected information in this document may be applicable for any CDC program that releases or shares data. However, this determination is left to the judgment of the CDC CIOs and their respective programs.

1. Background and Purpose

States[+] have a long-standing history of voluntarily reporting individually identifiable data to CDC on incident conditions or diseases that are of public health importance[1]. Recent developments in telecommunications and computerization have greatly enhanced the ability to compile and share such public health data. While the electronic exchange and accumulation of data on individual cases promises public health benefits, it has the potential to threaten individual privacy. The challenge is to balance the need for data protection with another competing interest of public health—the need to share data collected in the interest of public health as broadly as possible, with appropriate protections, with public health practitioners and with researchers conducting studies that have the potential to benefit public health. If such a balance is not achieved, potential data providers may choose to withhold data to protect it[2].

The U.S. State and Territorial Health Agencies operate within the authority of state-specific laws and regulations that control the collection and protection of individually

[+] Throughout this document, "States" should be understood to refer to both U.S. States and Territories.

identifiable data. Therefore, States are ultimately responsible for confidentiality protections, regardless of whether the data reside temporarily with a data steward, such as CDC, or reside within their own agencies.

Since the mid-1980s, the CDC and Council of State and Territorial Epidemiologists (CSTE) have been engaged in extensive discussions over issues related to re-release by CDC of data released by States to CDC. States wanted assurance that CDC programs would apply consistent principles and adhere to certain standards when releasing such data. Appendix A describes relevant events leading up to the establishment of the CDC-CSTE Intergovernmental Data Release Guidelines Working Group (DRGWG).

CDC has a responsibility to ensure that CDC programs protect the confidentiality of the State-provided data they have been entrusted with, and inform CSTE and data providers how the confidentiality of this data is being protected. In addition, CDC and CSTE have a shared responsibility to develop feasible guidelines for CDC programs that are consistent with State laws, regulations, and policies protecting confidentiality and that reflect state-of-the-art or best practices[3,4,5,6]. The *CDC-ATSDR Data Release Guidelines and Procedures for Re-release of State-Provided Data* are intended to: 1) complement existing Federal laws that govern data release and protect confidentiality (Appendices B1 and B2); 2) augment current CDC policies such as *The CDC Staff Manual on Confidentiality*[7] and the *NCHS Staff Manual on Confidentiality*[8]*;* and 3) provide an implementation guide for the newly published *CDC/ATSDR Policy on Releasing and Sharing Data*[9] (www.cdc.gov/od/foia/policies/sharing.htm), with respect to the re-release

of State-provided data. This report should be used by CDC programs when developing their program-specific data re-release procedures for State-provided data.

This report describes guidelines (minimum standards) for the development of program-specific data release plans and procedures for re-release of State-provided data by CDC. These guidelines are consistent with and expand upon the requirements listed in the *CDC/ATSDR Policy for Releasing and Sharing Data*[9]. A key principle of this report is the need for CDC programs to develop data agreements with State data providers before the data are received by CDC. The report also recognizes that in some CDC programs, States currently release data to CDC in the absence of explicit data agreements, and a process is necessary to develop these agreements even as data sharing continues. To facilitate this process, CDC programs will develop data re-release plans based on accepted data release practices and procedures as well as scientifically acceptable principles for confidentiality protection. Prior to finalization, draft CDC program-specific data release plans will be shared with data providers for their input. Formal agreement for each data re-release plan will be obtained from the data providers through an opt-in or opt-out "statement of response" from the data providers. For data providers deciding to opt-out of the data release plan, further negotiation with the data providers will be needed to customize the plan to meet State requirements.

Data collected by CDC, including data collected by States and provided to CDC, becomes Federal record once received by CDC, and is subject to Federal laws and rules governing data release and Federal records retention laws. These include, but are not

limited to, the Freedom of Information Act (FOIA), the Privacy Act of 1974, Confidentiality Assurances and Certificates of Confidentiality (see Appendix B for further details). These laws, which are highlighted in Appendices B1 (Federal Laws and Rules Governing Data Release) and B2 (overview of selected Federal laws), may provide CDC with the ability to protect certain types of data from public re-disclosure; they also may require the retention and/or disclosure of data in some circumstances. Data use agreements must conform to the requirements of these laws when applicable.

2. What Type of Data do the CDC-ATSDR Data Release Guidelines and Procedures Apply to?

The *CDC-ATSDR Data Release Guidelines and Procedures for Re-release of State-Provided Data* contain information that may be applicable for any CDC program that releases or shares data; however, it has been prepared specifically to address the policies and practices CDC programs establish for State-provided data shared with CDC that are not already covered by a written formal data re-release procedure at the time this report is finalized. State-provided data are defined, in this report, as population-based data intended to represent a complete count of cases (or a statistical sample of all cases in a given population) that are collected by U.S. State and Territorial Health Agencies related to the health or exposure status of individual U.S. residents, which fall under State legal authority for collection and protection of privacy and confidentiality, and that are reported to CDC by State health departments. The Guidelines and Procedures also apply

to unweighted microdata[10,11] (individual person records) from sample surveys administered by State Health Agencies which send CDC data containing personal identifiers or information about survey respondents that could potentially be used to identify survey respondents. Furthermore, these data release guidelines and procedures apply to State-provided surveillance and information data about events (such as a chemical spill or conflagration, etc.) only if data has been provided by the State in the format of individual person records associated with the event (see the glossary definition of "individually identifiable data," which mentions that both direct and indirect identifiers can potentially be used to establish individual identity).

Data from Indian Tribal nations that comes to CDC **indirectly** through State Health Agencies are covered by the guidelines and procedures in this report because these data are being directly reported to CDC under the authority of the State Health Agency. For CDC public health surveillance systems which are comprised of data from State Health Agencies and other data sources, such as the Vaccine Adverse Events Reporting System, one data release procedure should be developed and that procedure should not be in conflict with the data release guidelines for State-provided data.

Although most of the guidelines are applicable for any data that are released or shared, data **not** specifically covered by the *CDC-ATSDR Data Release Guidelines and Procedures for Re-Release of State-Provided Data* include, but are not limited to, the following:

- individually identifiable data that are not collected under U.S. State health agency authority, such as data that CDC collects directly, and data that are reported directly to CDC by Indian Tribal nations;
- data collected specifically for research or an outbreak investigation;
- data that are not individually identifiable or potentially identifiable;
- data systems that use public-use data compiled from another data system considered to be the primary data system.

3. Agreements with Data Providers

A. Introduction

This section and the next (Section 4) describe 16 guideline elements (the "Guidelines") that collectively constitute the *CDC-ATSDR Data Release Guidelines and Procedures for Re-release of State-Provided Data*. Each guideline represents a "minimum standard" for CDC programs to address when developing their program-specific data release procedures for re-release of State-provided data. CDC programs may wish to adopt more stringent standards than the minimum standard. The guidelines in Section 3 apply to all re-releases of State-provided data. The guidelines in Section 4 are grouped into three main categories: (1) guidelines representing administrative requirements that apply to all re-releases of State-provided data, (2) guidelines that apply to re-releases of State-provided data as public-use data, and (3) guidelines that apply to the re-release of State-provided data as restricted-access data.

Best practices statements accompany most of the guidelines. The best practices statements reflect applicable references and resources or selected examples of practices by CDC programs or other Federal or State programs that appear to be consistent with the guidelines. On occasion, more than one best practices standard is shown because they may represent viable alternative options for specific CDC programs. The best practices statements differ from the guidelines, in that the best practices statements are not minimum standards, but rather a listing of approaches that may merit more discussion and consideration within CDC. The best practices statements are meant to be descriptive rather than prescriptive. In other words, it is being left to the discretion of each CDC

program to consider for themselves whether it is appropriate to implement the listed best practices statements or implement another approach to meet the intent of the guideline. Because CDC data systems vary widely with respect to their content and format, it is important that both the guidelines and best practices allow for flexibility within the context of the principles espoused by the CDC policy.

The guidelines that apply to the re-release of restricted-access data using Data Sharing Agreements (DSAs) (see Section 4C) do not apply to re-release of restricted-access data through CDC-controlled data centers or by licensing non-CDC researchers to use certain data. At the time this report was developed, only NCHS and one program within the National Immunization Program were re-releasing data through a research data center and no CDC programs were re-releasing data through licensing agreements. Hence, until CDC implements a process for instituting these alternative mechanisms for data re-release more broadly within CDC, it is considered premature for this report to address the procedures needed for these alternative data re-release mechanisms. In the future, CDC and CSTE will need to develop additional implementation policies and procedures for data re-released through these mechanisms. For information about the NCHS Research Data Center, refer to www.cdc.gov/nchs/r&d/rdc.htm. Additional information about data centers[5,12] (controlled sites for accessing data) and data licensing[5,13] can be found elsewhere.

The cross-reference table included below lists the guidelines (see Sections 3 and 4A-4C) and procedures for implementing confidentiality protection (see Section 4D) that apply to

re-releases of microdata and tabular data (including pre-calculated tables) through online query systems or other formats.

Cross-Reference Table for Guidelines and Confidentiality Procedures that Apply to CDC's Re-release of Microdata or Tabular Data

Microdata	Tabular data
Guidelines 1-16; confidentiality protection procedures 1-6	Guidelines 1-4, 6, 16; confidentiality protection procedures 1, 2, 4-6

B. Interacting with State Data Providers to Facilitate Negotiation of a Data Re-Release Plan Proposed by CDC

Guideline 1: Develop Data Agreements with State Data Providers

A key principle of this guideline is the need for individual CDC programs to develop data agreements with State data providers before data are received by CDC. For CDC programs that are currently receiving data from States in the absence of written explicit data agreements with their State data providers, data sharing need not be interrupted as such agreements are being established. To facilitate the negotiation of the data agreement with State data providers, CDC programs will provide State data providers with a statement describing the *CDC/ATSDR Policy on Releasing and Sharing Data*[9] and the CDC program's proposed data re-release plan. The CDC program's proposed data re-release plan will be consistent with the *CDC/ATSDR Policy on Releasing and Sharing Data*[9] and will be based upon accepted data release practices and procedures (see Section 4A-C for guidelines on protecting and releasing data and see Section 4D for a description of confidentiality protection standards and procedures). The data re-release plan will also describe the content and format of data to be released as either non-identifiable public-

use data or identifiable/potentially identifiable restricted-access data (see also Guideline 2).

There may be some aspects of the data re-release plan that may need to be deferred until the data are reviewed by the CDC program. In this situation, a short-term (and most likely brief) data re-release plan could be drafted in collaboration with the State data providers prior to the time CDC receives the data and then a more comprehensive longer-term data re-release plan developed after the CDC program has reviewed the data. The short-term plan could simply remind the State data providers about any assurances and approvals that are in place to protect the data, if any, and inform them that no data, or very limited data, will be re-released before the longer-term plan is developed (see Figure 1 below for boilerplate language for a short-term data release plan and data agreement with data providers). The short-term initial data agreement with State data providers should specify the expected time the initial plan will remain in place as well as the number of months the CDC program anticipates it will take to complete the longer-term more comprehensive data re-release plan in collaboration with State data providers.

CDC programs will solicit the State data provider's input and formal agreement with the proposed data re-release plan by asking the data providers for a formal "statement of response," which at a minimum should include the State's decision to either "opt-in" or "opt-out" of the plan or components of the plan (e.g., a State may want to include their data in a CDC program's re-release of public-use data, but not restricted-access data). The CDC program may want to customize the proposed data re-release plan for State data

providers wishing to opt-out of the plan because it does not offer adequate protection for their State's data. For example, if feasible, the CDC program could withhold re-release of data below a specific sub-state geographic level for some States. While States have the option to refrain from providing data to CDC if the CDC plan provides less protection than State requirements, this should be a very rare occurrence that is only used as a last resort option by the State. The CDC program will review and, if necessary, update the data release plan periodically, and will notify States whenever a change in procedure is anticipated.

The provisions of the data re-release plan will be consistent with all applicable Federal and State laws and regulations. The CDC program will be responsible for determining that the agreements with data providers meet the requirements of the Federal laws and regulations under which it operates (see Appendix B1 and B2 for a summary of the Federal laws that apply to CDC's re-release of data). Similarly, the State data providers will be responsible for determining that the data agreement is consistent with all State laws and regulations under which it operates.

Figure 1. Boilerplate Language to be Used by CDC Programs for the Initial Data Agreement with State Data Providers, Prior to the First Data Submission to CDC for a New Surveillance System.

Data Agreement with State Data Providers

The **[*name of CDC program*]** is in the process of establishing the **[*insert name of surveillance system*]** in collaboration with U.S. States. Beginning **[*insert the date*]**, the **[insert name of CDC program]** will pilot test the reporting of data from data providers to the **[*name of the surveillance system's*]** data repository located at CDC. [***Insert the name of the CDC Program***] requests your State's participation in submitting data for this pilot test period, which is expected to last **[*insert number of months*]**. The primary purpose of data submitted during this period is to **[*insert reason, such as, test the system for receipt, evaluation of data quality, evaluate completeness of case ascertainment, etc*]**. In addition, this pilot will enable us to review the data for the purpose of developing a more comprehensive data re-release or data sharing plan.

The *CDC/ATSDR Policy on Releasing and Sharing Data* (www.cdc.gov/od/foia/policies/sharing.htm was developed to ensure that (1) CDC routinely provides data to its partners for appropriate public health purposes and (2) all data are released without restrictions or shared with particular parties with restrictions, as soon as feasible without compromising privacy concerns, Federal and State confidentiality concerns, proprietary interests, national security interests, or law enforcement activities. Data provided to CDC by State Health Departments are covered by this policy. This policy recognizes the importance of evaluating data quality and preparing appropriate documentation of the data, including instructions for non-CDC users on the appropriate use of the data, before re-releasing or sharing the data.

No data will be re-released or published from the data submitted during the pilot test period. However, we may share data back with the State data providers for the purpose of data quality review **[*or list alternative uses of the data*]**.

> ***OR a possible alternative to the above paragraph is:*** The only data planned for re-release during this period will be tabular data for CDC reports and presentations which will represent aggregated counts of **[*insert name of disease/condition*]** across all participating States by **[*period of time, such as month*]** for the purpose of **[*insert purpose*]**. A data disclaimer will accompany these data when they are included in CDC reports and presentations, informing data users of the test nature and provisional nature of the data.

By the end of the pilot test period on **[*insert date*]**, the **[*insert name of the CDC program*]** will have completed the development of a more comprehensive data re-release policy. The development of the longer-term and more comprehensive data re-release plan will be done in collaboration with State data providers, will result in the development of data agreements with each of the State data providers, as per CDC-ATSDR implementation guidelines for the *CDC/ATSDR Policy on Releasing and Sharing Data* (see Guidelines 1 and 2 in the *CDC-ATSDR Data Release Guidelines and Procedures for Re-Release of State-Provided Data*), and is expected to take no longer than **[*insert the number of months*]**.

The **[*insert name of CDC program*]** will keep you informed of the major activities we plan for developing the data re-release plan and State data agreements.

The data submitted to the **[*insert name of surveillance system*]** are covered by **[*insert the assurance or Federal law protecting the data, if applicable, such as a 308(d) Assurance of Confidentiality*]** which protects the data from **[*insert how the data are protected*]**. Attached is the statement of Assurance of Confidentiality for the **[*name of the surveillance system, and only include this sentence if it is applicable*]**.

Any requests the **[*insert name of the surveillance system*]** receives for data submitted during this pilot test period will be referred back to the State data providers. The more comprehensive data re-release plan will address the need for CDC to re-release data when requests are made for data submitted outside the pilot test period.

►**Best practices for Guideline 1**: The NCHS and NAPHSIS have an agreed upon data re-release agreement for vital statistics data.

The CDC AIDS Program asked States to select the level of sub-State geographic detail that specific variables in the AIDS Public Information Data Set (PIDS) could be tabulated for, by pre-defining a select group of options they could select from, such as health-district level; county-level; MSAs with 100,000 or more people; MSAs with 500,000 or more people, etc. This process enabled CDC to customize the AIDS data re-release procedure to meet the different requirements of individual States.

Guideline 2: Suggested Content of the Data Re-Release Plan

The content of the data re-release plan will vary according to the needs of the CDC program and data providers. The following are examples of topics the CDC program may want to include in the plan:

- Statement describing the *CDC/ATSDR Policy on Releasing and Sharing Data*[9];
- Statement indicating the re-release of data is consistent with Federal laws pertaining to the re-release of data;
- Data documentation descriptions and issues (see Appendix D in the *CDC/ATSDR Policy on Releasing and Sharing Data*) such as conditions under which the data were collected, the extent of the data's completeness and accuracy, potential limitations on the use of the data;
- Memoranda of Understanding, if any, that CDC has with other organizations that may involve use of the data.

For public-use data re-release (if any):

- Data elements to be re-released and special instructions on format;
- Nature of the confidentiality protection (see section 4D);
- Provisional and emergency data re-release plan (see Guidelines 6 and 16);
- Public Release Disclosure Statement (see Guideline 9);
- Schedule for data re-release;

For restricted-access data re-release (if any):

- Description of who can access restricted-access data files;
- Data elements to be re-released and any special instructions on format;
- Nature of the confidentiality protection (see Section 4D);

- Provisional and emergency data re-release plan (see Guidelines 6 and 16);
- Description of authentication procedures (see Guideline 10);
- Content of Data Sharing Agreements (see Guidelines 13, 14, 15);
- Method of monitoring compliance with terms of the Data Sharing Agreement (see Guideline 12).

4. Selected Procedures for Protecting and Releasing State-Provided Data

Using the following guidelines, CDC programs will develop a proposed data re-release plan, which will be negotiated with State data providers in order to obtain their input and formal agreement with the proposed plan (see Guideline 1 and Guideline 2).

A. Administrative Requirements for All Re-Releases of State-Provided Data

Guideline 3: Designate a Privacy Manager

Each CDC program that re-releases State-provided data will designate a Privacy Manager to clear all proposed releases, including data re-releases made within the provisions of an inter-agency Memorandum of Understanding (MOU). The designated program unit data steward can act as the Privacy Manager.

> ►**Best practices for Guideline 3**: Oversight for the Privacy Manager can be provided by a CIO data-release review board, which might report to the CIO Associate Director for Science (ADS), and might include the CIO Information Resources Manager (IRM) and relevant data stewards (see *CDC/ATSDR Policy on Releasing and Sharing Data*[9]).

Guideline 4: Train All Responsible Staff

The CDC program that re-releases State-provided data will ensure that staff responsible for data release will complete special training in confidentiality protection and disclosure risk assessment and control. *Note: Curricula for "special training" have yet to be determined, but Practice A (see Section 5 of this report) to support re-release of data addresses the need for CDC to itself develop or to contract with others to develop training curricula.*

> ►**Best practices for Guideline 4:** Training for staff involved in making decisions about data release, at the time of first employment and annually thereafter, as a training refresher, would be ideal. The supervisors of data release decision-makers should also receive this training. Until CDC makes training available, a recent text can serve as a useful reference.[3] In addition, the Federal Committee on Statistical Methodology (FCSM) Confidentiality and Data Access Committee (CDAC) has developed a one-day short course, entitled "Privacy, Confidentiality and the Protection of Health Data - A Statistical Perspective," which is available through special request (see www.fcsm.gov/cdac/index.html).

Guideline 5. Classify each Data Set as a Restricted-Access or a Public-Use Data Set (PUDS) and Define Criteria for Access to Restricted-Access Files

Each CDC program which re-releases State-provided data will classify each data set, whether developed for "planned release" or created in response to a data request, as either a PUDS or a "restricted-access" data set. A PUDS is comprised of data that have been modified to the extent needed to block breaches of confidentiality and prevent identity disclosure or disclosure of confidential information. There should be no restrictions on access to PUDS (see Guideline 8). However, a public release disclosure statement should accompany the PUDS (see Guideline 9).

A restricted-access data set includes either the full data set a State provides to CDC for program planning and evaluation (including data sets from etiologic studies), or a version of the full dataset that has been partially or substantially modified to minimize the likelihood of breaches of confidentiality. Each State needs to confirm, via the "statement of response" (see Guideline 1) whether it supports re-release of the State-provided data as a restricted-access data file. As the name implies, access to these data are restricted as follows:

- For CDC and State and local public health employees, permission for use of restricted-access data sets will be limited to those employees having official programmatic duties warranting access to these data.
- For other data requestors, permission for use of restricted-access data will be based upon a review of the stated purpose of the data request (the stated purpose of the data request should be consistent with the original purpose for data collection), an assessment of whether the requested data would be appropriate to

use for the intended purpose, and the need for using restricted-access data versus another type of available data set (e.g., a PUDS).

The *CDC/ATSDR Policy on Releasing and Sharing Data* states: "CDC strives to have data release policies that are fair to all users, regardless of their organizational affiliation."

CDC programs should develop a procedure describing the criteria for determining who can access non-PUDS (i.e., restricted-access) data and the party or parties within CDC who are responsible for making these decisions. Procedures for releasing restricted-access data include authenticating the requestor's identity (see Guideline 10) and the use of data sharing agreements (DSA) (see Guidelines 11, 13, 14, 15).

►Best practices for Guideline 5:
The *NCHS Policy on Micro-data Dissemination*[14] (www.cdc.gov/nchs/about/policy/policy.htm) states: "No individual–at NCHS or elsewhere–may claim entitlement to obtain or access identifiable data collected by NCHS by virtue of his or her employment. Access to identifiable data is not determined solely by employment status, organizational affiliation, or financial commitment. More important are the need for identifiable data, the use to which the data will be put, and the requestor's role and responsibility with respect to the data collection activity. Since any access to identifiable data poses risk, access to such data will be carefully evaluated and tracked after access is granted."

The CDC/NCHSTP assurance of confidentiality for HIV/AIDS data states "No CDC HIV/AIDS surveillance or research information that could be used to identify any individual or institution on whom a record is maintained, either directly or indirectly, will be made available to anyone for non-public health purposes. In particular, such information will not be disclosed to the public; to family members; to parties involved in civil,

criminal, or administrative litigation, or for commercial purposes; to agencies of the Federal, State, or local government. Data will only be released to the public, to other components of CDC, or to agencies of the Federal, State, or local government for public health purposes in accordance with the policies for data release established by the Council of State and Territorial Epidemiologists."

The 'minimum necessary' standard provision in the HIPAA Privacy Rule indicates that a "covered entity must make reasonable efforts to limit protected health information to the minimum necessary to accomplish the intended purpose of the use, disclosure, or request." Applying the intent of this standard to CDC data re-release procedures, only the minimum necessary data elements that can be justified by the data requestor to accomplish the proposed analysis should be re-released to data requestors, because additional variables may increase the disclosure risk potential of the data unnecessarily.

Guideline 6: Include Disclaimer with Re-Release of Provisional Data

'Provisional' or 'preliminary' data are thought to be close to final but subject to change as additional records are added to the dataset or updated information is obtained. The exact definition of 'provisional' or 'preliminary' varies by data system. Because provisional data may be subject to substantial change, it may not be appropriate for these data to be used for all purposes for which finalized data are used. The CDC program will inform State data providers of their procedure for any proposed re-release of provisional public-use and restricted-access data and States, in turn, will indicate their agreement or disagreement with the re-release procedure via their "statement of response" (see Guideline 1). A "Provisional Data Disclaimer" should accompany provisional data re-releases. It should encourage the data user to consider the provisional nature of the data before using it for decisions. In addition, the disclaimer should describe how the data are reviewed to ensure accuracy, and when the data are considered finalized.

►**Best practices for Guideline 6**: Data quality review is considered complete upon mutual agreement by the State and CDC, once the maximum achievable quality has been attained. Explicit criteria for data finalization should be documented and include adequate time for corrections from data providers and for case investigation to establish final case classification.

Guideline 7: Maintain Log of Data Sets Re-Released

The CDC program which re-releases State-provided data will maintain a log of data sets released.

►**Best practices for Guideline 7:** For PUDS data, the log of data sets released represents an inventory of the different PUDS data sets CDC has released. Ideally, such an inventory would be posted on the Internet, include a description of all State-provided PUDS public health surveillance data released by CDC, and be formatted so it can be queried by users. While the inventory is useful to inform interested parties as to the existence of these data sets, its primary function within CDC is to remind the CDC program that a potential exists for data users to combine related data sets in order to access more information than just a single data set would provide. Thus, the inventory should prompt CDC programs to perform disclosure risk assessment within the context of all previously released related data sets. Minimum data elements for the PUDS inventory could include the date the PUDS was first released, conditions or diseases included in the data set, variables and coding formats contained within the PUDS, and information on how to request the PUDS. If more than one PUDS data set is released by a single CDC program, the program should clarify how their PUDS data sets are different.

For restricted-access data, the primary purpose of the log of releases is to be able to track the status and terms of each data sharing agreement (DSA). This log is for internal CDC use by the CDC program responsible for tracking compliance with the terms of the DSA. Ideally, such a log would be posted on the CDC Intranet. The log of releases for restricted-access data sets should be audited to assess whether the person or persons granted access to restricted-access data are complying with the terms of their

DSA. Minimum elements for the restricted-access log may include name and affiliation of the person responsible for compliance with the terms of the DSA and information on how to contact them, names of all collaborators on the project (and information on how to contact them), the list of variables and coding formats released, the name of the conditions or diseases represented by the data, and a checklist of requirements the CDC program needs to verify as per the DSA (For example, if a pre-publication review of the report was required by the DSA, the date the data user sent the report to CDC for review and the date the review was completed and comments sent to the data user; or, the date the data set was returned to CDC or destroyed, etc.).

B. Re-release of State-Provided Data as Public-Use Data

CDC's re-release of data as a PUDS does not require the use of a DSA. When PUDS are created by CDC programs, they should be made available to all interested users, without restrictions.

Guideline 8: Planning for Release of Public-Use Data

Refer to the *CDC/ATSDR Policy on Releasing and Sharing Data*[9] for information about the release of data for public use. This Policy indicates "procedures for releasing public-use data should be consistent with the CDC's Public Health Information Network's functions and specifications[15]." The CDC/ATSDR Policy indicates each plan for release of public-use data should include the following:

- A procedure to ensure that confidential information is not disclosed.
- A procedure to ensure that data is released in a form that does not endanger national security or compromise law enforcement activities.

- Analysis plans and other documentation required by the Office of Management and Budget regulation on data quality[16].
- Instructions for non-CDC users on the appropriate use of the data.
- The date the data will be released, which should be as soon as possible after the data are collected, scrutinized for errors and validated. The release of these data should occur no more than one year after these activities are completed.
- The formats in which the data will be released (e.g., ASCII). For each format, give specifications (e.g., variable definitions) and information on standards for transmission.

The *CDC/ATSDR Policy on Releasing and Sharing Data*[9] also states CIOs may release data without restrictions for public use through the CDC Information Center or data may be shared through the CDC/ATSDR Scientific Data Repository and its data dissemination portal CDC WONDER (wonder.cdc.gov/welcome.html).

Guideline 9: Include Disclosure Statement with PUDS

At the time each PUDS is released or accessed, CDC programs will include a written statement about the following responsibilities users of public-use data have:

- A statement informing PUDS users of their responsibility to maintain confidentiality, including (per the *CDC/ATSDR Policy on Releasing and Sharing Data*[9]), "instructions that non-CDC data users must agree not to link data with other data sets…..[and]… instructions to report to the CDC ADS any inadvertent discovery of the identity of any person and to make no use of that discovery."

- A statement informing PUDS users of their responsibility not to imply or state, either in written or oral form, that interpretations based on the data are those of the original data sources (e.g., the U.S. States) or the CDC public health surveillance program that provided the data, unless the data user and data sources are formally collaborating on the proposed analysis.
- A statement informing users of their responsibility to acknowledge, in all reports based on these data, the original source of the data (e.g., the States that provided the data to CDC) as well as the name of the CDC public health surveillance program that re-released the data.

►**Best practices for Guideline 9:** CDC NCHS requests that PUDS users agree to:

- "Use the data in this dataset for statistical reporting and analysis only.
- Make no use of the identity of any person discovered inadvertently and advise the Director, NCHS, of any such discovery.
- Not link this dataset with individually identifiable data from other NCHS or non-NCHS datasets."

The data use agreement for "restricted-access public-use data" from the Agency for Healthcare Research and Quality (AHRQ), Healthcare Cost and Utilization Project (HCUP) and other similar data sets, includes the following statement: "I will make no statement nor permit others to make statements indicating or suggesting that interpretations drawn are those of data sources or AHRQ."

C. Re-release of Data as Restricted-Access Data

C1. Administrative Controls

Guideline 10: Authenticate the Identify of Data Requestors

When a request for re-release of State-provided restricted-access data is received by the CDC program, CDC staff will authenticate the identity of the requestor. Authentication methods can be paper-based or electronic. An electronic authentication protocol can use software access control, a physical device, or biometric scan.

►**Best practices for Guideline 10:** Following are examples.

- Written requests for access to State-provided data are required to be on letterhead stationery.
- Oral or email requests from a known individual are considered as needing no further verification.

An electronic authentication protocol can use one or more of the following methods.

- Software access control: Password or challenge phrase; Digital certificate
- Physical device: Hardware "token" (e.g., USB connection); Digital fob (auto-generated passnumber); SmartCard
- Biometric scan

The electronic process for identification-authentication can be linked to authorization "rights" which specify levels of access.

Guideline 11: All Requestors Wanting to Use Restricted-Access Data are Required to Sign a DSA

The CDC program should confirm, via the "statement of response" (see Guideline 1) that the State granted permission for the data to be re-released as a restricted-access data file.

In addition, prior to CDC's re-release of these data, all requestors must sign a DSA which governs the protection and use of these data. The CDC program which re-releases State-provided data will retain signed copies of the DSA. A DSA may be subsumed in a larger interagency Memorandum of Understanding (MOU).

►**Best practices for Guideline 11**: The 2002 modification to the HIPAA Privacy Rule permits release of a "limited data set" as long as a there is a written data use agreement. This HIPAA standard is consistent with the use of "restricted-access" data and DSAs described within the guidelines section of this report.

Guideline 12: Monitor User Compliance with DSAs

The CDC program which re-releases State-provided data will monitor compliance with the terms of the DSA.

►**Best practices for Guideline 12:** A passive approach to compliance monitoring is acceptable, as long as the following criteria are met:

- The data user is informed of the penalty for not complying with the terms of the DSA. (The intent of the penalty is to deter breaches in compliance.)
- The data user is fully informed about their responsibilities in using the data.
- The data steward institutes a process for logging compliance problems they become aware of.
- The CDC program takes appropriate actions to resolve any identified problems and implements (if possible) procedures to avoid similar types of problems in the future.

Examples of active compliance monitoring methods include:

- Pre-publication review of reports, articles, graphs, maps, or tables.
- Prior review of presentations based on the dataset.
- Annual letters sent by data stewards to confirm whether the data requestor's use of the dataset has been completed and

whether the data requestor has taken steps to either destroy or return the dataset.

Pre-publication or pre-presentation review can include both privacy protection as well as accuracy of scientific inferences.

C2. Development of Data Sharing Agreements for Restricted-Access Data

Guideline 13: Requirements for a Standard DSA for Restricted-Access Data

Projects with a 308(d) assurance of confidentiality may have additional standards and requirements beyond those described below and those requirements may vary by CDC program, even within the same CDC CIO. CIOs should consult with the Management Analysis and Services Office (MASO) Privacy Officer to determine if any additional standards apply to a project with a 308(d) protection, in order to ensure compliance with the law. If the requestor plans to use the data for research and plans to obtain, or has obtained, approval for the study from an officially sanctioned Institutional Review Board (IRB), the DSA may be abbreviated as follows: the approved protocol or the notice of IRB approval may be substituted for portions of the DSA, as deemed appropriate by the CDC program. Otherwise, every standard DSA must include the following elements, which include not only the required criteria listed in the *CDC-ATSDR Policy on Releasing and Sharing Data*[9] for "special-use agreements," but also includes additional criteria:

- A description of the use to which the data will be put, and limitations on usage of data. The data requestor's description of their intended use of the data should provide evidence to the CDC program that there is a legitimate public health purpose that justifies the use of the data. The data user should also demonstrate their need for restricted-access data versus other available data, such as a PUDS.
- Information on any laws pertaining to the DSA.
- The names of every person who will have access to the data and specification of procedures for extending the provisions of the DSA to named collaborators (e.g., requiring signed confidentiality pledges, etc).
- The name of the person primarily responsible for care of the released data and compliance with the terms of the agreement.
- A list of mechanisms for preservation of confidentiality. These mechanisms should include both limitations on access (i.e., specified staff only) and technical security practices (such as encryption).
- A list of restrictions on releasing analytic results.
- A clearly stated prohibition on any attempt to link the dataset with any other dataset without prior CDC permission (see Guideline 14).
- A clearly stated prohibition on the further release of data to other parties without prior CDC permission (see Guideline 15). Requests from legal authorities (such as under conditions of a declared public health emergency) or FOIA requests must be referred by the data user to the CDC data steward.
- A stated requirement for the data user to notify the CDC ADS if any individual person represented in the dataset is inadvertently identified during approved usage.

- A stated limitation of the right of access to data based on the role of an individual, with a stated requirement to return or destroy the dataset when the requestor changes positions in the agency or leaves the agency.
- A stated requirement to return or destroy the dataset and all derived files when use is completed (the term "use" in this sense may include a specified plan for re-analyzing the data after the initial analysis is completed).
- A statement informing users of their responsibility not to imply or state, either in written or oral form, that interpretations based on the data are those of the original data sources (for example, the U.S. States) or the CDC program that provided the data, unless the data user and data providers are formally collaborating on the proposed analysis.
- A statement informing users of their responsibility to acknowledge, in all reports based on these data, the original source of the data (for example, the States that initially provided the data) as well as the name of the CDC program that re-released the data to the user (see best practices for an example).

- A description of the penalty that may be imposed on the data user for breaching the terms of the DSA.
- Provisions that govern emergency requests for identifiable or otherwise confidential data (See also Guideline 16).
- For DSAs with CDC staff acting as a data user, the following are to be included in the DSA:
    - A statement indicating that if the CDC data user plans to publish a report that singles out specific States or Cities in the discussion section of the report, the

CDC program should send a courtesy message and copy of the report to the State, before the expected publication date. This requirement does not apply in situations where a table in the report includes all 50 States. (It is understood that the CDC programs will use their judgment in determining when to contact the State. The intent of this requirement is to help ensure the State has time to prepare a response or reaction to news media or to other inquiries arising from publication of the report.)

- A statement describing how the CDC program intends to monitor compliance with the terms of the DSA, such as pre-publication review of reports, presentations, maps, graphs, etc (see Guideline 12).

The *CDC/ATSDR Policy on Releasing and Sharing Data* indicates that data that cannot be publicly shared may be shared with restrictions with public health partners. The CDC/ATSDR Policy indicates "Restrictions can be imposed because of legal constraints or because releasing the data would risk: (1) disclosing proprietary or confidential information; or (2) compromising national security or law enforcement efforts...For restricted data, special data sharing agreements must be developed." The CDC/ATSDR Policy also indicates, data shared with restrictions can be shared (1) using CDC-controlled data centers (such as NCHS has developed); (2) by licensing non-CDC researchers to use certain data, if CDC chooses to consider this licensing option in the future; or (3) through a special-use agreement. This guideline and the Standard DSA both apply to "restricted-access data" shared outside of CDC-controlled data centers.

►**Best practices for Guideline 13:** The South Carolina Office of Research and Statistics application for release of data includes the

following statement to data requestors: "All releases of data must contain the following statement: NOTICE: THIS INFORMATION IS FROM THE RECORDS OF THE OFFICE OF RESEARCH AND STATISTICS, BUDGET AND CONTROL BOARD, SOUTH CAROLINA. OUR AUTHORIZATION TO RELEASE THIS INFORMATION DOES NOT IMPLY ENDORSEMENT OF THE STUDY OR ITS FINDINGS."

CDC/EPO provides the following information regarding acknowledgement to users of data from the National Notifiable Diseases Surveillance System (NNDSS):
"We request that any published material derived from NNDSS data acknowledge 1) the U.S. State and Territorial Health Departments that collect the data from a range of case ascertainment sources (e.g., health-care providers, hospitals, laboratories) and report these data to CDC; 2) CDC's National Notifiable Diseases Surveillance System; and 3) the Surveillance Systems Branch, Division of Public Health Surveillance and Informatics, Epidemiology Program Office (responsible for preparing and aggregating State-provided NNDSS data for dissemination)."

Guideline 14: Include Addendum to the Standard DSA When a Data Requestor Plans to Link Restricted-Access Data to Other Data

Restricted-access data include identifiable or potentially identifiable data. The data set derived from linking a restricted-access data set to another data file may be even more individually identifiable than the original data. Thus, there is a need to ensure disclosure review takes into account the variables included after the linkage.

The CDC program interested in re-releasing restricted-access data for a linked data analysis should obtain, or require the data requestor to obtain, written permission from the data providers (the States contributing the data) for the proposed linked analysis. Prior to seeking permission from the data providers, the data requestor should complete

and sign the standard DSA (see Guidelines 11, 13) and, in addition, attach an addendum to the DSA which includes the following information, so the CDC program and data providers can determine whether to approve the data request: 1) source and description of the data file to which the restricted-access data will be linked; 2) written description of the variables and coding formats to be included in the final linked file; and 3) description of the data requestor's plan for conducting additional disclosure review to ensure that variables contributed in the linking process do not lead to re-identification of the individual described in the original data file.

►**Best practices for Guideline 14**: Examples of special procedures for linked analyses are:

- "separation of duties", where no single individual is able to conduct all of the steps in the linkage process.
- post-linkage de-identification. Standard de-identification methods should be used, such as numerator and denominator cell aggregation or suppression rules.

GAO Report #GAO-01126SP titled "RECORD LINKAGE AND PRIVACY: Issues in Creating New Federal Research and Statistical Information[21]" may serve as a useful reference of techniques which can be employed to ensure privacy protection for data linkages.

Linkage can be conducted inside a CDC-controlled data center. This guideline applies to "modified" data shared outside of such centers. Data shared outside CDC-controlled data centers are partially or substantially modified, as needed, in order to minimize the likelihood of breaches of confidentiality.

Guideline 15: Include Addendum to the Standard DSA When a Data Requestor Plans Further Data Releases from Restricted-Access Data to Other Parties

Restricted-access data can be released further to other parties as de-identified data, in one of two ways: 1) by creating a de-identified data set (i.e., a PUDS) from the restricted-access data set and then disseminating data from the PUDS, or 2) by generating de-identified data directly from the restricted-access data set (where the restricted-access data file, by definition, includes identifiable or potentially identifiable data).

If the requestor plans to release de-identified data generated directly from a restricted-access data file (situation #2 described above), a procedure must be implemented to ensure that the generated data have been de-identified appropriately. In this situation, the data requestor is required to submit an addendum to the standard DSA (See Guidelines 11, 13) describing the procedures that will be implemented to audit the results of the data generated for further release to other parties.

If the requestor plans to release de-identified data that are generated from a de-identified data set created from the restricted-access data set (situation #1 described above), an addendum to the standard DSA is not required.

►**Best practices for Guideline 15:** If the requestor plans release via an online interactive query system which generates tables from restricted data, an automated algorithm for de-identification must act prior to release, and its results must be checked by an automated audit protocol. *An automated audit protocol assesses disclosure risk against pre-determined thresholds. (See Procedure 4.)*

C3. Emergency Requests for Data

Guideline 16: Emergency Requests for Data

A request for restricted-access data that poses an important public health question to address a public health emergency, that has a high likelihood of yielding results important for understanding, controlling, or responding to a public health emergency, and is consistent with CDC program data re-release procedures of the pertinent CDC program, should undergo expedited review and be fast-tracked for processing. As with other data re-releases, the *CDC/ATSDR Policy for Releasing and Sharing Data* indicates released data should not compromise privacy concerns, Federal and State confidentiality concerns, proprietary interests, national security interests, or law enforcement activities.

Because health conditions related to terrorism events may be very rare events (have low incidence in the population) and may be highly publicized in the media (as evidenced by the anthrax contamination of the U.S. mail in 2001), numerator and denominator cell size aggregation and suppression rules may not be appropriate for these types of events. Data re-release for such events may require the development of special data sharing agreements. Emergency requests for restricted-access data that violate existing CDC program data release procedures should undergo expedited review by the CIO data-

release review board in order to assess whether the request should be denied on the basis that it violates the CDC program data release procedures or should be approved with special restrictions and constraints outlined in a special data sharing agreement. Referral to the CIO data-release review board may help the board prospectively develop criteria for how to address similar data requests in the future. A special data sharing agreement will need to be developed and tailored to meet the needs of the particular emergency situation.

D. Confidentiality Protection

A. Current Standards for De-Identifying Data Sets and Performing Disclosure Review Assessment

It is the intention of this report to encourage CDC-ATSDR Programs to re-release data as much as possible through the use of PUDS. The creation of a PUDS will require CDC programs to go through a data de-identification process since PUDS data do not contain individually identifiable information. In the opinion of DRGWG, the current standards used to help de-identify data sets are as follows: 1) the Federal Committee on Statistical Methodology's Statistical (FCSM) Working Paper 22 on statistical disclosure methodology,[17] which includes an educational primer on statistical disclosure limitation; 2) the *Checklist on Disclosure Potential of Proposed Data Releases*, developed by the Interagency Confidentiality and Data Access Committee, FCSM[18]; and 3) The Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule[19]. However, the HIPAA Privacy Rule's "safe harbor" method of de-identification is deemed inadequate protection for public health agencies to use for de-identifying data sets because public

health agencies collect many more demographic variables than do covered health entities. In addition, public health agencies are expected to have a much higher level of sophistication with regard to disclosure review than entities covered by HIPAA.

Other useful materials on confidentiality protection have been developed and will be useful to those involved in data protection and re-release, including a book published in 2001 titled *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*[3]; a 1993 book titled *Private Lives and Public Policies, Confidentiality and Accessibility of Government Statistics*[4]; the Institute of Medicine's report *Protecting Data Privacy in Health Services Research*[20]*;* and the Government Accounting Office report on record linkage and privacy[21]. In addition, the FCSM's Confidentiality and Data Access Committee has prepared a document titled "Identifiability in Microdata Files"[22] that is instructive in terms of helping one think through the factors which could increase the re-identification potential of a data set.

Prior to CDC's re-release of data, the data should be subjected to a disclosure review analysis that has included the variables earmarked for re-release. Other issues included in the CDC staff manual on confidentiality[7], within the section pertaining to avoiding inadvertent disclosures in published data, that may help in the analysis of the disclosure risk potential of data include:

1) Types of Disclosure
    a) exact versus approximate,
    b) probability-based versus certainty disclosures,

c) internal versus external disclosures,

2) Issues relevant to evaluating a disclosure problem

a) sampling ratio the data are based on,

b) existence of errors or imputations in the data,

c) incompleteness of reporting, and

d) sensitivity of the data.

3) Measures to avoid disclosure

a) aggregate data to eliminate the particular cells that would otherwise produce a disclosure, or

b) use primary and secondary cell suppression edits.

B. Procedures for Implementing Confidentiality Protection

The following is a list of procedures for confidentiality protection. One or more of these procedures may be appropriate for a given data system. It is not the intention of this report to imply that all these methods need to be used for each data system. CDC programs may wish to seek expert consultation on the issue of confidentiality protection and disclosure risk assessment and control before finalizing a data re-release plan (see Practice E (in Section 5 of this report).

Procedure 1. Limit Disclosure of Potential Identifiers

The CDC programs that re-release State-provided data will delete, as necessary, all information judged to be of potential use in making individuals to whom they pertain

"identifiable" (see "individually identifiable data" in Glossary). Particular attention will be devoted to information that can lead directly to an individual or their family, such as:

- Name
- Street address
- Social Security Number
- Medical record number
- Telephone number

Other information is extremely useful in narrowing the possibilities that information may refer to a particular individual. Those providing low-level geographic detail and precise timing of certain events are particularly important.

Geographic Information

- Zip code (9-, 5-, and even 3-digit)
- Census tract
- City/town
- County

Timing of events

- Exact date of birth (year-month-day)
- Exact date of event (year-month-day)
- Month and year of birth (year-month)
- Month and year of event (year-month)

Information concerning location (e.g., geography) and timing of events, along with details concerning the following types of information can serve to reveal individual identity:

- Occupation (e.g. 3- or higher digit codes)
- Education (e.g. single years)
- Race/ethnicity (e.g. Aleutian, Filipino v. Asian)
- Income (as a continuous, un-topcoded variable)
- Medical condition/diagnosis (e.g. detailed site specific cancers) and cause of death (e.g. 3- or higher digit ICD code)

There could be other program-specific variables that may be used to help disclose individual identity.

Prior to re-release, these items must be thoroughly reviewed for their potential for personal re-identification. *It is not the intent of these guidelines to imply that all of these fields must always be deleted.* Under appropriate conditions and with proper safeguards, such items may be released. It is not the intent of these guidelines to recommend a single strategy for the limitation of disclosure risk. It is recognized that both programmatic and statistical considerations may come into play in deciding appropriate protections for data release. For example, an alternative to field (variable) deletion is data value recoding (see Procedure 2).

Prior re-releases of data should be considered in a cumulative fashion. As stated in the document *Identifiability in Microdata Files*[22] "Each of these files may be considered safe to release by themselves. However, there may be enough information in the two files combined to effect a re-identification."

►**Best practices for Procedure 1:** Refer to the Federal Committee on Statistical Methodology's Statistical Working Paper #22: Report on Statistical Disclosure Limitation Methodology[11] for a description of common methods used to protect microdata and tabular data (www.fcsm.gov/working-papers/wp22.html). Additional guidance is available in the Federal Committee on Statistical Methodology's Confidentiality and Data Access Committee Checklist on Disclosure Potential of Proposed Data Releases[6] (www.fcsm.gov/committees/cdac/index.html). Additional guidance as well as bibliographical materials on statistical disclosure limitation methodology can be found at this website as well as that of the Committee on Privacy and Confidentiality of the American Statistical Association (users.erols.com/dewolf/pchome.htm).

The Health Insurance Portability and Accountability Act[19] (HIPAA) 2001 Privacy Rule's safe harbor method of de-identification (which requires covered entities to remove all of a list of 18 enumerated identifiers and have no actual knowledge that the information that remains could be used alone or in combination to identify an individual who is a subject of the information) is deemed inadequate protection for pubic health agencies to use for de-identifying data sets because public health agencies collect many more demographic variables than do covered entities, according to the viewpoint of DRGWG representatives. In addition, public health agencies are expected to have a much higher level of sophistication with regard to disclosure review than covered entities.

The HIPAA Privacy Rule de-identification standard that is consistent with this guideline requires "a person with appropriate knowledge and experience applying generally acceptable statistical and scientific principles and methods for rendering information not individually identifiable makes and documents a determination that there is a very small risk that the information could be used by

others to identify a subject of the information" (45 CFR Part 164.514(b)(1)(i)).

The HIPAA Privacy Rule was modified in 2002. The 2002 modification to the HIPAA Privacy Rule permits release of data that is not fully de-identified through a "limited data set," as long as the data are governed by a data use agreement which provides sufficient privacy and confidentiality protection for the data. For the limited data set, all direct identifiers must be removed from the data file. However, dates such as birth date, date of death, and dates of admission or discharge could be released in the data file, but only if this information is needed for the purpose of the release. In addition, the limited data set may include 5-digit zip code or any other geographic subdivisions, such as State, county, city, precinct and their equivalent geocodes, except for street address. (See August 14, 2002 Federal Register Notice updating 45 CFR Parts 160 and 164 of the HIPAA Privacy Rule (CFR Part 164.514(b)(1)(i)), at www.hhs.gov/ocr/hipaa/privrulepd.pdf .

The Privacy Rule clarifies standards for the creation of de-identified data sets and "limited data sets." The limited data set concept is consistent with the concept of "restricted-access" data that is discussed in this report.

Procedure 2. Aggregate Data Values

The CDC programs which re-release State-provided data, either as "microdata" files (e.g., data files or records on an individual person) or as tabular data, will recode fields as needed in order to aggregate data values. Common methods include:

- collapsing continuous/interval data (e.g., age; date of occurrence) into broad categories;
- collapsing ordinal data (e.g., location geography) into broader categories;
- grouping nominal data (e.g., diagnosis) into broad categories;
- truncating variables (e.g., name).
- top-coding or bottom-coding variables.

►**Best practices for Procedure 2:** Refer to the Federal Committee on Statistical Methodology's *Statistical Working Paper #22: Report on Statistical Disclosure Limitation Methodology*[11] for a description of common methods used to protect microdata and tabular data, available at www.fcsm.gov/working-papers/wp22.html.

Procedure 3. Limit the Number of Records or the Number of Fields

The CDC program which re-releases State-provided data as "microdata" files will consider the applicability of other methods to protect microdata releases from disclosure risk. Common methods include:

- including data from only a sample of the full dataset;
- limiting the number of variables in the released file. For example, consider (1) releasing only those variables essential to analysis; and (2) limiting the number of contextual or ecological variables--information that describes a geographic area, such as where a case-patient resides--because this type of information can lead to the identification of that area.

►**Best practices for Procedure 3:** Refer to the Federal Committee on Statistical Methodology's *Statistical Working Paper #22: Report on Statistical Disclosure Limitation Methodology*[11] for a description of common methods used to protect microdata, available at fcsm.gov/working-papers/wp22.html.

Procedure 4. Use Numerator Rules for Data Aggregation or Suppression

The CDC program which re-releases State-provided data as tabular data or microdata will use numerator rules (cell size) to either

- guide selection of groupings of aggregated data values; or, if aggregation is insufficient,

- suppress release of certain cells in a table.

There is no single numeric threshold for cell aggregation that is appropriate for all data in tabular or microdata format. Selection of an appropriate threshold level is guided by multiple factors, including:

- sensitivity of the data (subject matter);
- format of the data (e.g., whether the data are continuous or categorical);
- level of detail in the data, especially the level of geographic detail;
- likelihood that a specific record in a database may represent a unique person in a small population;
- population or subgroup denominator size, as well as the numerator size.

►**Best practices for Procedure 4:** The following are some examples of aggregation rules (see also Healthy People 2010 Statistical Notes, Number 24, published by NCHS in July 2002, accessible at www.cdc.gov/nchs/products/pubs/pubd/hp2k/statnt/30-21.htm.

The CDC staff manual on confidentiality[7] generally advises a minimum numerator cell size of "3". Thus, numerator cell size counts of "1" or "2" are not generally advised, with some notable exceptions.

The CDC staff manual on confidentiality[7] indicates the following exceptions to minimum numerator cell size rule of "3":

a) "It has been a longstanding tradition in the field of morbidity or mortality [and vital] statistics not to suppress small frequency cells in the tabulation and presentation of data. For example, it has been considered important to know that there were two deaths from rabies in Rio Arriba County, N. Mex., in a given year, or that there were only one infant death and two fetal deaths in Aitkin County, Minn. These types of exceptions to general CDC practices in other programs are followed because they have been

accepted traditionally and because they rarely, if ever, reveal information about individuals that is not known socially."

b) "Tables may show simple *counts* of numbers of persons, even though the number in a cell is only '1' or '2' provided the classifying data are not judged to be sensitive in the context of the table..."

The Washington State guideline states: "If the count of cases or events in a cell is less than three, the data analyst needs to consider whether a breach of confidentiality is likely. A count of no events in the cell is clearly no threat to confidentiality, but a count of one or two events may be." (See www.doh.wa.gov/Data/Guidelines/SmallNumbers.htm.)

A rule which combines numerators and denominators will meet this standard, such as a rule that data are not released if the population is less than a certain size and the number of events in a cell is less than a certain size. For example, Missouri uses a complex combined rule[23]: "A table is not reported if a table cell subtracted from the number of total events of the same data file for the same characteristics yields a small number (less than 10)."

CDC staff should review guidelines for avoiding inadvertent disclosures in published data, in the CDC or NCHS Staff Manual on Confidentiality[7,8]. The following information about special guidelines to avoid disclosures in published data has been abstracted from the CDC staff manual on confidentiality[7] [Please note that at the time this report was being drafted, the CDC and NCHS staff manuals on confidentiality were being revised. The comments inserted below in square brackets reflect draft proposed changes to the NCHS manual on confidentiality, which is anticipated to be published in 2004.]:

Special Guidelines for Avoiding Disclosures:

A. In no table should all cases of any line or column be found in a single cell.

B. In no case should the total figure for a line or column of a cross-tabulation be less than three. [In the proposed updated NCHS staff manual on confidentiality, this statement is revised as follows: In no case should the total figure for a line or column of a cross-tabulation be less than 5 unwieghted sample cases.]

C. In no case should a quantity figure be based upon fewer than three sample cases. [In proposed updated NCHS staff manual on confidentiality, this statement is revised as

follows: In no case should a quantity figure be based upon fewer than five sample cases.]

D. In no case should a quantity figure be published if one case contributes more than 60 percent of the amount. [In the proposed updated NCHS staff manual on confidentiality, this statement is revised statement as follows: In no case should a quantity figure be published if one case contributes a disproportionate amount to the total. A minimum percentage figure should be adopted for this purpose and this figure should not be publicly released.]

E. In no case should data on an identifiable case, nor any of the kinds of data listed in the preceding items A-D, be derivable through subtraction or other calculation from the combination of tables published on a given study.

F. Data published by CDC should never permit disclosure when used in combination with other known data.

The appropriate level of cell suppression may vary based on the geographic level of detail being presented. For example, at the national level, there may be no reason to suppress any data, but as the denominator size decreases, such as at the State, county, or smaller geographic level, cell suppression may need to be employed to prevent inadvertent re-identification of the person described in the data base.

At the Agency for Healthcare Research and Quality (AHRQ) Healthcare Cost & Utilization Project (HCUP) interactive query website[24], AHRQ suppresses "Values based on 30 or fewer discharges" from tabular results of State HCUP tables. (See www.ahrq.gov/data/hcup/hcupnet.htm.)

When primary cell suppression is used, it is generally accepted that 1) complementary cell suppression will need to be employed to avoid back-calculation by subtraction; and 2) if tabular data are suppressed using automated algorithms on a large number of related tables or via web-based interactive query systems generating tables directly from raw data, that procedures for post-suppression auditing should be employed. However, precise methods for automated suppression and auditing, that are easily implemented, still need to be developed. Therefore, at this time, it is not possible to implement this "best practice," but it should be considered for future implementation.

Procedure 5. Use Denominator Rules for Data Aggregation or Suppression

The CDC programs which re-release State-provided data as tabular data or microdata will use denominator rules (population size) to either:

- guide selection of groupings of aggregated data values; or, if aggregation is insufficient,
- suppress release of certain cells in a table.

►**Best practices for Procedure 5:** A rule which combines numerators and denominators will meet this standard, such as a rule that addresses the relationship between the size of the numerator and the size of the denominator.

An approach commonly used for microdata is that data are not released if the total population from which the data are drawn is less than a certain size, based on the premise of a size sufficiently large that no subcell of the variables contained in the data would be expected to be smaller than a certain size.

In considering the population size in tabulated data, guidelines employed by the Washington State Department of Health state that "Generally, tabular data based on denominators greater than 300 persons per cell present minimal risk for individual identification. ... Caution should be exercised by the analyst if the population size is between 100 and 300, and extreme caution is warranted when the population is less than 100." (See www.doh.wa.gov/Data/Guidelines/SmallNumbers.htm.)

As indicated in Procedure 4 above, in the case of tabulated data Missouri focuses on the number of people in the population with the characteristic indicated in a given numerator. If the difference between the two is less than a minimum number, this information is not be published. On the other hand, in this scheme a very small cell number could be published if the denominator were large enough. (See www.cdc.gov/epo/dphsi/AI/confiles/day1/Land1.ppt).

With regard to *total* population size, the statistical literature contains a great deal of discussion of appropriate minimal sizes and there is clearly variation depending upon information content

and special considerations. The most general statement is found in the Federal Committee on Statistical Methodology's Statistical Policy Working Paper 2 where it is stated "Geographic information must be restricted beyond the point where an individual user could be familiar with a significant proportion of the universe, but whether that point comes at 25,000, 250,000 or 1 million will depend on the detail in the file and other restrictions imposed." (See www.fcsm.gov/working-papers/sw2.html p 28).

The Federal Committee on Statistical Methodology's Confidentiality and Data Access Committee Checklist on Disclosure Potential of Proposed Data Releases[6] calls for "a minimum of 100,000 persons in the sampled area … [else] provide rationale." (See fcsm.gov/committees/cdac/checklist_799.doc).

The NCHS version of this checklist contains the following language:
"Generally one has to balance the level of survey detail against the level of geography. The greater the amount of detail, the more risk is entailed for lower levels of geography. Similarly, with very high levels of geography, greater detail may be made available.

General Rule: All geographic areas that are identified must have a minimum of 100,000 persons in the sampled area (according to latest Census or Census estimate).

Caution: *the figure of 100,000 is not without some risk.* For certain target populations the members of which are be found infrequently in a population, a higher number may be desired."

Procedure 6. Refrain From Using Techniques that Distort Data for Privacy Protection

The CDC programs which re-release State-provided data will refrain from distorting data (either altering data values or omitting records from the dataset) unless this approach is employed as a last resort and is absolutely necessary for the purpose of privacy protection.

►**Best practices for Procedure 6:** Examples of distorting data include adding statistical noise, data swapping, blanking and imputing for randomly selected records, and blurring data (replacing a reported value by an average value). Current methods of perturbative "statistical disclosure control" are not optimal because one cannot prospectively assess how the distortion of the data will affect the results of an analysis. This is a problem particularly if recommendations for public health action are made based on the analysis of distorted data. At the discretion of the CIO, when a data requestor plans to make public health recommendations based on analysis of altered data, the CDC program holding that data may offer to confirm the requestor's findings by performing a re-analysis on un-altered data. Alternatively, researchers should be provided the opportunity to obtain restricted-access data under a data sharing agreement, or conduct re-analysis themselves in a CDC-controlled research data center.

See also Practice G which mentions that a new form of disclosure limitation, entitled "controlled tabular adjustment," is being researched; software is being developed to implement this new method. It was beyond the scope of this report to assess which data distortion methods minimize the magnitude of data distortion.

5. Practices to Support Re-Release of Data

A. Development of Curricula for "Special Training"

CDC should determine the specific content of the curricula for "special training" that data stewards and other members of the CIO data-release review board (or other group having oversight responsibility for data re-releases) should have (as per Guideline 4: Training). This should include training in confidentiality protection and disclosure risk assessment and control (as per Guideline 4: Training) at a minimum, but it may also be beneficial to consider including training in other issues, such as enterprise-wide security standards for public health surveillance, or any other subjects CDC identifies if it conducts a staff

training needs assessment. The specific issues to cover within the above subjects and the frequency of staff training are left to the discretion of CDC. However, CDC should require the staff receiving training to pass a post-training test with a grade that demonstrates comprehension of the training material. CDC may wish to develop training modules on the above subjects itself or contract with experts in the above subjects to develop training materials and resources.

B. Development of a Data Set Inventory to Facilitate Disclosure Risk Assessment

Disclosure risk assessment is performed to estimate, either qualitatively or quantitatively, the probability that a data set poses a high or low risk of re-identification in terms of the information it contains about individuals and the status of their health. Since the risk of re-identification may be increased if a one data set can be linked to an another data set, CDC data stewards should be aware of the data sets other CDC components release that could have the potential for linkage with data sets their own programs release. To help facilitate this aspect of disclosure risk assessment, CDC should compile and regularly update an inventory of data sets that have already been released and that are eligible for release in the near future. The inventory should be posted on the CDC Intranet in a browsable (or queriable) format and should include, for each data set, the name of the CDC program releasing the data set, data steward contact information, and data set documentation (see the "Documentation" section of the *CDC/ATSDR Policy on Releasing and Sharing Data*[9] for recommended categories of information data set

documentation should include), including the names of diseases or conditions about which the data are tabulated , and the variables and coding formats used.

C. Development of Instructions Data Stewards can Use to Create PUDS for Data Re-Releases, including FOIA Requests

CDC should consider the feasibility of developing specific instructions for creating PUDS that CDC program data stewards could use. If feasible, CDC should develop or contract with others to create PUDS development instructions. Having PUDS development instructions would also help data stewards to create de-identified data sets for Freedom of Information (FOIA) data requests, since the CDC FOIA Office does not currently provide any standardized criteria to CDC data stewards to help them in creating de-identified data sets for release in response to FOIA requests.

D. Evaluations to Assess Whether a Breach of Confidentiality has Occurred

This report and the *CDC/ATSDR Policy on Releasing and Sharing Data* states that potential confidentiality breaches should be reported to the CDC ADS. This passive approach to ascertaining potential confidentiality breaches assumes that CDC will receive such reports. CDC should consider the feasibility of taking a more active approach to identifying confidentiality breaches. If feasible to do so, CDC may wish to itself develop or contract with others to develop standard criteria for CDC programs to use in conducting active approaches.

E. Consultation with Experts on Confidentiality Protection and Disclosure Risk Assessment

This report is not intended to be (and cannot be) a comprehensive resource on confidentiality protection or disclosure risk assessment and control. CDC programs may need to consult with experts on these issues as they develop program-specific data re-release procedures that are consistent with guidelines and procedures in this report. CDC may wish to generally offer their programs the services of specified experts on these issues. In addition, CDC may wish to develop an interest group forum on confidentiality and data release patterned after the Confidentiality and Data Access Committee (CDAC). CDAC, an interest group of the Office of Management and Budget's Federal Committee on Statistical Methodology (fcsm.gov/committees/cdac/cdac.html), was formed because staff members of statistical agencies who worked in the "confidentiality area" expressed a need to have a forum where they could communicate among themselves and exchange ideas.

F. Establish a CDC Intranet Site Where Materials Referenced in this Report, from Various Websites, Are Archived

This report cites various materials that are currently posted on the Internet pertaining to "training" issues, such as data release policies, confidentiality protection, or disclosure review assessment that would be useful to preserve for use with this report. To preserve the future availability of these materials, CDC should consider creating an Intranet site where these materials are archived.

G. Need for Continuing Discussions of Methods for Privacy Protection, Disclosure Risk, and Other Issues

The adoption of these guidelines should not deter CDC, CSTE, and others from future discussions of new methods for privacy protection and disclosure risk assessment and review. The guidelines and procedures are not "written in stone" and should be considered subject to change as new information and methods become available. For example, at the time this report was developed, the Working Group was not aware of any CDC programs using secondary (complementary) suppression based on methodologic principles, to avoid back-calculation (by subtraction) of the content of table cells that have undergone primary suppression. In addition, the Working Group was not aware of any algorithms being used by health agencies for automated suppression and auditing of one or more related tables. As noted in the best practices section for confidentiality protection Procedure #4 (see Section 4D, Procedure 4), precise and easy-to-implement methods for automated suppression and auditing of tables still need to be developed. Because use of web-based interactive query systems for data dissemination is increasing, future discussions should focus on the development and use of automated suppression and auditing methodologies for use by CDC programs.

In 2002, the NCHS Office of Research and Methodology (ORM) sponsored the development of methodologic software for complementary cell suppression (and controlled rounding and controlled random perturbation) in two-way statistical tables of counts or magnitudes. CDC may wish to explore these automated methods for

complementary cell suppression which are statistically-based in lieu of using ad-hoc or manual approaches.

The NCHS ORM also noted that research is underway for a new form of disclosure limitation in tables called "controlled tabular adjustment," which is not limited by the dimensionality or complexity of tables. While Procedure 6 indicates CDC programs should refrain from using disclosure limitation techniques that distort data, it may be useful for CDC to assess or contract with others to assess which of the various perturbative statistical disclosure control methods minimize the magnitude of data distortion, for potential use by public health systems.

Since CDC program data stewards will most likely be the primary staff responsible for writing program-specific procedures, under the direction of the CIO data-release review board or another oversight mechanism, CDC may wish to consider establishing a cross-CIO forum for data stewards to brainstorm and share suggestions for implementing the *CDC-ATSDR Data Release Guidelines and Procedures for Re-Release of State-Provided Data*.

Various issues were raised during the review of this document that the DRGWG did not address, but that merit future discussion, including the following concerns and issues that were expressed:

- CDC may fail to inform data users of potential weaknesses or peculiarities of the data when it releases the data, while the State may have provided such information to the user.
- CDC may re-release data selectively so that wrong inferences would be drawn from it, while the State would have released a different dataset in response to the same request.
- CDC may fail to inform the States of re-releases it has made of data that came from that State.
- CDC may re-release data containing uncorrected errors, while the State may have released the same data with corrections. More generally, data re-released by CDC may not be exactly the same as the corresponding dataset held by the State, and the user would not be informed of this possibility.
- CDC collects and uses data from various State agencies other than State health agencies, such as labor departments, environmental departments, and agriculture departments. These other sources of data were not included within the scope of the *CDC-ATSDR Data Release Guidelines and Procedures for Re-release of State-Provided Data* because of the way the Data Release Guidelines Working Group was constituted. Future discussions within CDC should focus on whether the implementation guidelines for the *CDC-ATSDR Policy for Releasing and Sharing Data* in this report should also apply to data from State agencies other than State health agencies.

- State-provided data that comes to CDC from other Federal agencies and not directly from State health agencies are not considered within the scope of the *CDC-ATSDR Data Release Guidelines and Procedures for Re-release of State-Provided Data*. Each request for this type of data may need to be handled differently, depending on the specifics of the situation, such as whether an MOU or other written agreement exists between the two Federal agencies or other statutory protections exist that define how external requests for data will be handled. In addition, for FOIA requests, an assessment may be needed to identify whether any FOIA exemptions may apply or whether the situation warrants a referral by the FOIA Officer to the Federal agency which was the source of the data being sent to CDC. Future discussions within CDC should focus on efficient processes for handling these types of data requests.
- Interest within CDC seems to be growing in terms of re-releasing data on the Internet in the form of online queries. The *CDC-ATSDR Data Release Guidelines and Procedures for Re-release of State-Provided Data* includes information relevant to the application of specific guidelines or procedures to re-release of data through an online interactive query system (e.g., see best practices for Guideline 15 and Procedure 4). In addition, the CDC CIOs may do work that should be broadly shared, such as the *Guide for Public Health Agencies Developing, Adopting, or Purchasing Interactive Web-based Data Dissemination Systems*[25] which was developed under a CDC contract with ORM Macro, Inc, and is posted along with other material

relevant to online queries on the CDC/EPO Division of Public Health Surveillance and Informatics Capacity Building Web Page (See www.cdc.gov/epo/dphsi/asb/orcmacro.htm) .

6. Implementation Steps

Proposed Deadline and Steps for CDC's Implementation of the Guidelines and Procedures

CDC and ATSDR programs having surveillance systems that fall within the scope of the CDC-ATSDR data release guidelines and procedures should examine their data re-release practices as of the effective date of the *CDC-ATSDR Data Release Guidelines and Procedures for Re-release of State-Provided Data* to see if they meet the minimum standards. If their procedures do not meet the minimum standards, CDC programs should have two years to revise their procedures to bring them into conformance, unless an appeal for an extension is requested from and granted by the CDC ADS.

CDC CIOs should be responsible for ensuring that the guidelines and procedures are implemented either through the establishment of a CIO data-release review board (see the *CDC/ATSDR Policy on Releasing and Sharing Data*[9]), which might report to the CIO ADS and might include the CIO Information Resources Manager and relevant data stewards, or CIOs might wish to implement the policy using an alternative oversight mechanism.

The OPS Announcement that is distributed after the *CDC-ATSDR Data Release Guidelines and Procedures for Re-release of State-Provided Data* are cleared by CDC should indicate the guidelines apply to data shared between the States and CDC that are not already covered by a formal written data re-release procedure. CDC programs should examine their practices, as of the distribution date, to see if they meet the minimum standards in guidelines, and if not, to revise them to bring them into conformance with the minimum standard guidelines. CDC program data re-release procedures should be forwarded to the CDC ADS Office where an assessment of conformance with the *CDC-ATSDR Data Release Guidelines and Procedures for Re-release of State-Provided Data* will be done.

7. Feedback to CSTE

Feedback to CSTE Regarding CDC's Implementation of the Guidelines

CDC should advise CSTE on at least an annual basis during the implementation phase of these guidelines, on the status of completion of implementation. In addition, CDC should communicate with CSTE regularly regarding the results of evaluations conducted after the guidelines and procedures have been implemented by CDC programs, particularly if the results of the evaluations indicate that a revision of the guidelines is warranted.

8. DRGWG Members

CDC Members

CIO	**Primary Representative**	**Alternate Representative**

ATSDR	Maureen Orr	Shannon Rossiter
EPO	Ruth Ann Jajosky (DRGWG co-chair)	
NCBDDD	Leslie O'Leary	
NCCDPHP	Nedra Whitehead	
NCEH	Sharunda Buchanan	
NCHS	Julie Kowaleski	Alvan Zarate
NCHSTP	Pat Sweeney	Sam Costa
NCID	Katherine Robinson	
NCIPC	vacant as of March 2002	
NIOSH	William Eschenbacher	Robert Castellan
NIP	Gail Horlick	
OD	Kenya Ford	
PHPPO	vacant as of April 22, 2002	

Former CDC Members

ATSDR	Wendy Kaye
NCID	Bob Pinner
NCIPC	John Horan
NCCDPHP	Luann Rhodes, Mary Hutton
PHPPO	William Yasnoff

CSTE Members

Affiliation	**Representative**

Washington State Department of Health	Steven C. Macdonald (DRGWG co-chair)
Iowa Department of Health	Patricia Quinlisk
Virginia Department of Health	Jennifer Haussler

Former CSTE members

Washington State Department of Health	Jac Davies
Kansas Department of Health and Environment	Gianfranco Pezzino

9. Acknowledgements

The DRGWG members wish to thank the following people: Aun Lor in CDC/EPO who worked with Betsey Dunaway, CDC Confidentiality and Privacy Officer in the Management Analysis and Services Office (MASO), to develop the document included in Appendix B2 of this report which summarizes the Freedom of Information Act, The Privacy Act of 1987, and Confidentiality Assurances (308(d)) and Certificates of Confidentiality (301(d)); Betsey Dunaway for the information and insights she shared with the DRGWG; Donna Stroup, Acting CDC ADS and former chair of the Excellence in Science sub-committee on data release and sharing, and the members of her sub-committee who drafted the *CDC/ATSDR Policy on Releasing and Sharing Data*[9]; Deborah Tress and Paula Kocher, CDC Office of General Counsel; Jennifer Madans, NCHS ADS; Richard Hopkins, Acting Director of the Division of Public Health Surveillance and Informatics, EPO; and lastly, a special thanks is due to the CDC, CSTE, and State Health Department leadership and supervisory staff for recognizing the

importance of the work undertaken by the DRGWG and for allowing their staff to dedicate time and resources to this effort.

10. Glossary

Audit trail: The maintenance of information, in a logbook or database, pertaining the request for and release and use of individually identifiable data.

Authentication: The process by which the identity of a person requesting access to individually identifiable data (restricted-access data) is verified.

Automated audit protocol: In terms of disclosure risk control, the use of linear programming to identify complementary cell suppressions for a primary cell suppression and to audit the proposed cell suppression pattern to see if it provides the required level of protection. Research seems to indicate that linear programming methodologies provide good but not optimal results. For this reason, it is not enough to just perform an automated audit for secondary cell suppression. The result of the algorithm needs to be checked to see if it is successful.

Bottom-coding: A technique used to mask microdata that involves creating categories for data values that are below a certain level. This method differs from aggregation, in that all data values below a certain threshold are grouped; other values in the field may or may not be grouped. See *top-coding*.

Cell suppression: One of the most commonly used ways of protecting sensitive cells in tabular data. It is obvious that in a row with a suppressed sensitive cell, at least one additional cell must be suppressed, or the value in the sensitive cell could be calculated exactly by subtraction from the marginal total. The same is true for the column which contains a suppressed cell. For this reason, certain other cells must also be suppressed. The suppression of a sensitive cell is termed a primary cell suppression. Suppression of other cells to prevent one from calculating the value in the sensitive cell is termed complementary (or secondary) cell suppression.

Computational disclosure control: The process by which data values are aggregated to increase the granularity of specific variables, such as grouping age into age groups prior to data release.

Confidentiality: The treatment of information that an individual or institution has disclosed in a relationship of trust, with the expectation that it will not be divulged to others in ways that are inconsistent with the understanding of the original disclosure. It encompasses access to and disclosure of information in accordance with requirements of law and/or official policy [26].

Confidentiality breach: **"**An unauthorized release of identifiable or confidential data or information, which may result from a security failure, intentional inappropriate behavior, human error, or natural disaster. A breach of confidentiality may or may not result in harm to one or more individuals." (Source: Washington State Department of Health "Guidelines for Working With Small Numbers, Glossary" (accessible at www.doh.wa.gov/Data/Guidelines/SmallNumbers.htm, accessed on April 11, 2002).

Data access: A general term referring to situations involving either data release or data sharing.

Data dissemination: Any mechanism by which data are made available to users. It includes mechanisms whereby data are released to users as well as mechanisms whereby data are made available without being released.

Data release: Dissemination of data either in a public-use file or as a result of an ad hoc request which results in the data steward no longer controlling the use of the data. Data may be released in a variety of formats including, but not limited to, tables, microdata (person records), or online query systems.

Data sharing: Granting certain individuals or organizations access to data that contain individually identifiable information with the understanding that individually identifiable or potentially identifiable data cannot be re-released further unless a special data sharing agreement governs the use and re-release of the data and is agreed upon by CDC and the data provider(s).

Data sharing agreement (DSA): A mechanism by which a data requestor and CDC program can define the terms of data access that can be granted to requestors.

Data steward: This person is responsible for the management, processing, documentation, integrity, and security of information in a data system. Data stewards can assume the responsibilities that Privacy Managers may have, such as the developing and implementing data confidentiality procedures, and being responsible for clearing responses to data requests for a surveillance system. (See the definition of *Privacy Manager* in this glossary.)

Disclosure: In this report, *disclosure* refers to the unauthorized public disclosure of information about a person, about which data have been collected. A disclosure may occur as a result of a confidentiality breach. The definition of disclosure used in the HIPAA Privacy Rule is different from the definition in this report. For purposes of HIPAA, *disclosure* means the release, transfer, provision of access to, or divulging in any other manner of information outside the entity holding the information.

Disclosure (risk) assessment: A systematic review of a data file conducted to determine if any of the proposed contents present an unacceptable risk of individual disclosure. Disclosure risk assessment and control are usually conducted to prepare a public-use data set, and they can also be conducted when preparing a data set that is potentially linkable to another released data set.

Disclosure (risk) control (also referred to as *disclosure limitation* or *disclosure protection*): The application of measures to reduce the possibility of identifying an individual through the characteristics available in a data file. Disclosure risk assessment and control are usually conducted in order to prepare a public-use data set, and they can

also be conducted when preparing a data set that is potentially linkable to another released data set. Disclosure control includes steps taken to modify or suppress information that might identify an individual directly or indirectly before the data are made available to others for analysis (see page 3 of the Doyle book[3]). Perturbative disclosure control methods distort (alter) the data before it is released while nonperturbative methods do not alter the data, but instead partially suppress or reduce the detail of the original data set (see page 112 of the Doyle book[3]).

Individually identifiable data: Data or information which can be used to establish individual identity, either directly, using items such as name, address, or unique identifying number, or indirectly by linking data about a case-individual with other information that uniquely identifies them.

Microdata: A data file containing information in which each record provides information at the unit of data collection (e.g., individual persons, events, households, or establishments).

Penalties: Penalties for a breach of confidentiality can range from imposing fines or a prison sentence to disciplinary action, barring an individual from receiving data in the future, or termination of employment or contract. Penalties can be established to differentiate willful from inadvertent disclosure and they can be tailored to the type of party responsible for the breach of confidentiality--an employee, contractor, or external data requestor.

Population-based data: A complete count of cases occurring within a given population or a statistical sample of all cases occurring within a given population.

Predecisional exemption: The Freedom of Information Act's "predecisional" exception is explained in 5 U.S.C. § 552(b)(5) as "inter-agency or intra-agency memorandums or letters which would not be available by law to a party other than an agency in litigation with the agency." The term "predecisional" has been traditionally defined by the courts as meaning "antecedent to the adoption of an agency policy." [27]

Privacy: The right of individuals to hold information about themselves in secret, free from the knowledge of others[28].

Privacy Manager: A person who develops and implements a data system's confidentiality policy and is responsible for clearing responses to data requests for a surveillance system. A data steward may act as a privacy manager. (See definition of *Data steward* in the glossary).

Proprietary: Produced or collected in such a way that exclusive rights may apply.

Provisional or preliminary data: These data are thought to be close to final but subject to change as additional records are added to the dataset or updated information is

obtained. The exact definition of 'provisional' and 'preliminary' varies by data system (see Guideline 6).

Public health emergency: An occurrence or imminent threat of an adverse health event caused by epidemic of pandemic disease, infectious agent, biologic or chemical toxin, environmental disaster, or any agent that poses a real and substantial risk for a significant number of human fatalities or cases of permanent or long-term disability.

Public-Use data: Data available to any requestor. Public-use data are sometimes referred to as "de-identified data" and include public-use data sets (PUDS), tables, or other data formats, with all individually identifiable data or information removed, and with the remaining fields modified or suppressed so as to reduce disclosure risk as much as reasonably possible and such that it is not possible to create any tables that violate the numerator or denominator cell size rules for a given surveillance system.

Public-Use Data Set (PUDS): See p*ublic-use data*.

Re-release: For purposes of this report, CDC's re-release of data that the States initially provided to CDC through their data release procedures. Data can be re-released to data users who are internal or external to CDC in various formats, including CDC reports, publications, graphs, tables, maps, presentations, and data files.

Restricted-access: Allowing the use of an Agency's microdata under controlled conditions. Restricted-access may mean allowing the use of an agency's data only to those who sign a formal data sharing agreement or permitting access to the data only at CDC-controlled research data centers, where CDC exercises direct supervision of the data use in order to protect confidentiality.

Security: The mechanisms (administrative, technical, physical) by which privacy and confidentiality policies are implemented in computer and telecommunication systems.

Top-coding: A technique used to mask microdata that involves creating categories for data values that exceed a certain level. This method differs from aggregation, in that all data values above a certain threshold are grouped; other values in a field may or may not be grouped. See *bottom-coding*.

11. Appendices

Appendix A: History leading to the establishment of the CDC-CSTE Intergovernmental Data Release Guidelines Working Group

CDC and CSTE have been engaged in extensive discussions over issues related to CDC's re-release of State-provided data. A June 1996 CSTE document, entitled *Data Release Guidelines of the Council of State & Territorial Epidemiologists for the National Public Health Surveillance System*, refers to these earlier discussions: "In 1985, CDC and CSTE jointly negotiated a policy for the release of data from CDC's notifiable disease surveillance system to facilitate its use for public health, while preserving the confidentiality of the data." Similarly, the 1996 agreement cites "A revised data release policy exclusively for AIDS was approved by the CSTE Executive Committee in 1995." The 1996 agreement was intended to update and broaden the 1985 policy, and it was anticipated by CSTE that it would subsequently be implemented by CDC in all of its CIOs. Although implementation was successful in a few program units (most notably the National Notifiable Diseases Surveillance System [NNDSS] in the Epidemiology Program Office), CDC did not implement the agreement in most other CIOs.

In 1999, the Surveillance Systems Branch in the Epidemiology Program Office (EPO), CDC prepared draft data release procedures for the NNDSS, which were based on the June 1996 *Data Release Guidelines of the Council of State and Territorial Epidemiologists for the National Public Health Surveillance System*. The draft 1999 NNDSS procedures were not intended to conflict with or revise CSTE's June 1996 guidelines. Instead, the draft 1999 NNDSS procedures were intended to clarify specific issues related to the release of NNDSS data and to formalize the 1996 CSTE guidelines. Surveillance Systems Branch staff sought consultation from CSTE's Surveillance Committee regarding implementation issues and problems they had experienced since 1996 related to ambiguities in CSTE's June 1996 data release guidelines. During this consultation, the Surveillance Systems Branch staff also asked CSTE of their interest in expanding the cell size suppression rules described in the June 1996 data release guidelines, from a consideration of not only the number of persons in the numerator of a table cell but also to include a consideration of the number of persons in the denominator of a table cell. The CSTE Surveillance Committee expressed interest in working with the Surveillance Systems Branch to finalize the draft 1999 NNDSS procedures, including preparing an expansion of the cell size suppression rules. They also expressed interest in learning about other CDC program practices related to re-release of State-provided data, including data CDC obtains through cooperative agreements. During the consultation, several CSTE members recollected that their expectation back in 1996 was that CSTE's June 1996 data release guidelines would be implemented CDC-wide; however,

subsequent inquiry in year 2000 into this issue by the now defunct CDC Surveillance Coordination Group indicated that other CDC CIOs were unaware of the CSTE's June 1996 data release guidelines. Shortly after the initial Surveillance Coordination Group inquiry, CSTE informally informed the Surveillance Systems Branch of their intent to send a letter to the CDC Director about the need to implement uniform data release practice at CDC for the re-release of State-provided data held at CDC. In addition, CSTE indicated their desire to address the uniformity of data release issue within CDC would temporarily delay the finalization of the 1999 NNDSS data release procedures.

In a letter mailed in May, 2000 to the CDC Director, the CSTE President expressed interest in establishing a joint CDC-CSTE working group to review CDC program-level data re-release practices pertaining to State-provided data and to develop and implement uniform CDC-wide data re-release guidelines for State-provided data. In response, Barbara Holloway, Acting Director of EPO, sent a letter in June 2000 to the CSTE President indicating support for such a collaboration and stating the CDC Health Information and Surveillance Systems Board (HISSB) Surveillance Coordination Group would play a critical role in addressing these issues. The CDC Surveillance Coordination Group chairman established the CDC-CSTE Intergovernmental Data Release Guidelines Working Group in February 2001. Then, in the fall of 2001, the CDC Excellence in Science Committee's data release and data sharing sub-committee assumed administrative oversight of the DRGWG, after CDC dissolved the HISSB and its working groups, including the Surveillance Coordination Group.

Appendix B1. Federal Laws and Rules Governing Data Release

CDC's practices and policies regarding data release are based on the framework of Federal Laws governing the maintenance and public disclosure of Federal records, the protection of key public priorities such as privacy, proprietary information, and national security, and obligations arising out of litigation or other compulsory processes. The purpose of this section is to provide an overview of the major Federal Laws that may affect the release of State data or that may compel disclosure of State data. The applicability and implications of these laws will vary depending on the nature of the data and other circumstances. These variations will not be analyzed in detail here. Questions by data stewards about the applicability of legal requirements should be directed to appropriate legal counsel.

Data collected by CDC, including data collected by States and provided to CDC, generally become a Federal record once received by CDC, and are subject to Federal laws and rules governing data release and Federal records retention laws. These include but are not limited to, the Freedom of Information Act (FOIA), the Privacy Act of 1974, Confidentiality Assurances, and Certificates of Confidentiality. These legal authorities are highlighted in Appendix B2 (overview of selected Federal laws). Guidance on the application of the HIPAA Privacy Rule to public health is provided in a recently published *Morbidity and Mortality Weekly Report* supplement[29] (www.cdc.gov/mmwr/pdf/wk/mmsu5201.pdf). These Federal laws may provide CDC with the ability to protect certain types of data from public re-disclosure; they also may require the retention and/or disclosure of data in some circumstances. Data use agreements must conform to the requirements of these laws when applicable.

The Federal Records Act (44 U.S.C. Chapter 33, 36 CFR Chapter 12, Subchapter B, Records Management) prescribes how and for how long Federal records are to be maintained, when they may be destroyed, and when they may or must be archived. In general, original data received by CDC in the form of paper or electronic data are kept for a standard period of ten years, and then archived or destroyed in accordance with the approved CDC Records Control Schedule (Item 2-47, Research Working Papers B-231). CDC's retention of this information does not negate the requirement that organizations (e.g., State agencies) follow records retention guidelines required under law and by local and State governments. In addition, requirements imposed by States requiring CDC to destroy or return data and records may conflict with CDC's record retention obligations.

The Freedom of Information Act (FOIA) (www.cdc.gov/od/foia/foi.htm) generally provides that, upon written request from any person, a Federal agency must release any agency record unless that record falls within one of nine exemptions. Since, as stated above, State-based data become a Federal record in CDC's possession, such records are subject to disclosure in response to a FOIA request. However, several of these FOIA exemptions may be available to protect some aspects of State data from public disclosures in response to a FOIA request (see Table 1 and Appendix B2).

The Privacy Act of 1974 (5 U.S.C. 552a) prohibits agency disclosure of any record maintained in a Federal system of records when the primary method by which the data will be retrieved is by name, social security number, or other identifying particular, such as thumb print, except pursuant to a written request, or with the prior written consent of, the individual to whom the record pertains. The creation of a Federal system of records is announced in the *Federal Register*. While the Privacy Act is generally protective, it contains several exceptions to the Privacy Act which permit disclosure without a subject's consent, including disclosure of the record for a routine use. A routine use is a disclosure which is compatible with the purpose for which the record was collected, and which is included in the system notice. A complete list of exceptions is found within Title 5 U.S.C. 552(a) (see Appendix B2).

In addition to the *Privacy Act of 1974*, which contains criminal penalties for the unauthorized disclosure of protected information, the *Trade Secrets Act* (Table 1) makes it unlawful for any officer or employee of the United States or of any Federal department or agency to publish, divulge, disclose, or make known in any manner not authorized by law any information gained through the course of Federal employment that concerns or relates to "trade secrets, processes, operations, style of work, or apparatus, or to the identity, confidential statistical data, amount or source of any income, profits, or losses, or expenditures of any person, firm, partnership, corporation, or association…" Information acquired through the course of official duties, through an investigation or examination, or seen in a document, is covered by the Trade Secrets Act. Some types of data CDC receives from States may fall under this act, making the unauthorized disclosure potentially a criminal offense. Under the Trade Secrets Act (18 U.S.C. Section 1905), a person "…shall be fined not more than $1,000, or imprisoned not more than one year, or both; and shall be removed from office or employment." Under the *Privacy Act of 1974*, Subsection 552a(i) (1), a person who willfully discloses information to another person who is not entitled to receive it "…shall be guilty of a misdemeanor and fined not more than $5,000." The "Related Statutory Authorities" section of the Standards of Ethical Conduct for Employees of the Executive Branch (5 CFR 2635.902; see www.usoge.gov/pages/laws_regs_fedreg_stats/oge_regs/5cfr2635.html) cites the prohibition against disclosure of proprietary information and certain other information of a confidential nature which is contained in the Trade Secrets Act (18 U.S.C. Section 1905).

Special Confidentiality Protections. A limited number of CDC projects that collect highly sensitive identifiable information have received approval for formal confidentiality protection under Sections 301(d) or 308(d) of the Public Health Service (PHS) Act (42 U.S.C. Sections 241(d) and 242(m)(d) (see Appendix B). In addition, data collected by the National Center for Health Statistics (NCHS) is covered by section 308(d). Data collected under a project with a 308(d) assurance of confidentiality may not be released in identifiable form without the consent of the individual or entity that supplied the information. CDC has historically been rigorous in its approval process for the discretionary use of this authority in order to prevent misuse of the protections as a mechanism for refusing to share data.

Data collected under a 301(d) Certificate of Confidentiality may be protected from disclosure to persons not connected to the research. In addition, such researchers "…may not be compelled in any Federal, State, or local civil criminal, administrative, legislative, or other proceedings to identify such individuals."

HIPAA Privacy Rule (45 CFR Parts 160 and 164): CDC is not a covered entity under the HIPAA Privacy Rule, so the Privacy Rule does not apply to CDC and its handling of data provided by the States. Data recipients at CDC should be aware that some State health departments are covered entities under the Privacy Rule, and may require CDC to produce certain documentation related to their (the State's) obligations under the Rule. For example, States may request verification of identity, statutory authority, and a representation that the information requested by CDC is the minimum necessary for the particular activity. Also, if a State that is a covered entity provides a limited data set (as defined by the HIPAA Privacy Rule) to CDC and CDC enters into a data use agreement, CDC must comply with the terms of the data use agreement. Since CDC is still not a covered entity, it would not be subject to enforcement action or penalties under the Rule.

The HIPAA Privacy Rule also includes a standard related to the de-identification of data. Since CDC is not a covered entity this standard is not applicable to CDC's re-release of data. There may, however, be instances when it would be appropriate for CDC and other public health authorities to use the de-identification standard in the Privacy Rule to create a de-identified data set even if the Privacy Rule does not apply.

National Security. In some circumstances, State data disclosed to CDC may be relevant to national security. CDC may need to evaluate any risk to national security posed by a release of this data. CDC maintains classified information in accordance with Executive Order 12958. (See CDC Manual Guide No. CDC-5: Information Resources Management—Policy on Classified Material, rev 2002.) The Department of Health and Human Services (DHHS) was recently given authority to classify information in accordance with this order under 66 F.R. 64347 (2001). Determination of classification is made by the Secretary of DHHS. The need to protect sensitive but **unclassified** information from inappropriate disclosure is carefully balanced, on a case by case basis, with the benefits that result from the open and efficient exchange of scientific, technical, and like information. Release of such information to the public is made in accordance with FOIA.

Legal Proceedings. In some instances, data collected by CDC, including data collected by States and sent to CDC, may be implicated in some type of lawsuit or administrative proceeding. The disclosure of such data may be sought through voluntary disclosure or compulsory process. Generally, CDC uses available legal mechanisms to protect the confidentiality of identifiable data, and to protect other public interests described previously. Generally, CDC has been successful in such instances. (see for example, Farnsworth v. Proctor and Gamble Company, et al, 101 F.R.D. 355 (U.S. Dist. 1984). However, it is important to keep in mind that the ability to protect confidential or identifiable information in litigation or other legal proceedings depends on a variety of factors such as whether the data has special confidentiality protection, the type of court or

forum, whether it is a Federal or State proceeding, whether the United States is a party, and the particular laws at issue.

The *human subjects Common Rule* (Table 1) applies to all applications and proposals for research involving human subjects to be conducted, supported, or subject to regulation by a Federal department or agency. One of the many requirements of the Common Rule includes the provision, when appropriate, for the privacy of subjects to be protected (45 CFR Part 46, Section 46.111).

At the time this report was drafted, it was not possible to include a summary about how the Confidential Information Protection and Statistical Efficiency provisions of the E-Government Act of 2002 will impact CDC data systems. The Office of Management and Budget is expected to issue definitive guidance on this issue at a later date.

Table 1. Federal Laws and Rules Applicable to CDC's Release of Data

I. Federal Records Act
44 U.S.C. Chapter 33, 36 CFR Chapter 12, Subchapter B, Records Management

II. Freedom of Information Act (FOIA)
5 U.S.C. 552
www.usdoj.gov/foia/foiastat.htm
www.cdc.gov/od/foia/foi.htm
www.usdoj.gov/foia/04_3.html

III. Privacy Act
5 U.S.C. 552a
www.usdoj.gov/foia/privstat.htm

IV. Trade Secrets Act
18 U.S.C. Section 1905

V. Standards of Ethical Conduct for Employees of the Executive Branch
5 CFR 2635.902
www.usoge.gov/pages/laws_regs_fedreg_stats/oge_regs/5cfr2635.html

VI. Special Confidentiality Protections

Assurance of Confidentiality
Section 308(d) of the Public Health Service Act
42 U.S.C. 242(m)(d)

Certificate of Confidentiality
Section 301(d) of the Public Health Service Act
42 U.S.C. 241(d)

VII. DHHS Authority to Classify National Security
66 F.R. 64347 (2001)

VIII. Human Subjects Common Rule
45 C.F.R. Part 46, Section 46.111
ohrp.osophs.dhhs.gov/humansubjects/guidance/45cfr46.htm#46.111

IX. HIPAA Privacy Rule
45 CFR Parts 160 and 164

Appendix B2: Overview of the Freedom of Information Act, the Privacy Act, Confidentiality Assurances (308(d)), and Certificates of Confidentiality (301(d))

Introduction

Under the **Freedom of Information Act** (FOIA) (Title 5 United States Code §552 or 5 USC 552), **all** Federal agency records are subject to disclosure unless covered (in whole or part) by one (or more) of nine exemptions. The **Privacy Act** applies only to records in "systems of records*" [defined in 5 USC §552a(a)(5)], from which information is retrieved by an individual's name or other identifier. Such records are subject to release to individuals asking for their own records, to other requestors with the signed consent of the named individual, or to other requestors without a subject's consent under limited conditions specified in the Privacy Act. Also, Sections 301(d) (42 USC 241(d)-**Certificates of Confidentiality**) or 308(d) (42 USC 242m(d)-**Confidentiality Assurances**) of the Public Health Service Act, provide additional protection for the identities of individuals or institutions within records.

***Definition** - The term "system of records" means a group of any records under the control of any Federal agency from which information is retrieved by the name of the individual or by some identifying number, symbol, or other identifying particular assigned to the individual.

Freedom of Information Act (Title 5 USC 552)

The Freedom of Information Act (FOIA) provides that, upon receiving a written request from any person, a Federal agency must release any requested agency record unless that record falls within one of the nine FOIA exemptions. FOIA applies to only Federal agencies, and covers only records in the possession and control of those agencies except in certain narrow instances involving grantee-held data.

- Exemptions from FOIA (Title 5 USC 552b)

1. protects from disclosure national security information concerning national defense or foreign policy, provided that it has been properly classified pursuant to Executive Order 12,958.
2. related solely to the internal personnel rules and practices of an agency;
   (a) internal matters of a relatively trivial nature;
   (b) more substantial internal matters, the disclosure of which would risk circumvention of a legal requirement.
3. specifically exempted from disclosure by statute (other than section 552b of this title), provided that such statute:

(a) requires that the matter be withheld from the public in such a manner as to leave no discretion on the issue; or
(b) establishes particular criteria for withholding or refers to particular types of matters to be withheld;

4. trade secrets and commercial or financial information obtained from a person that is privileged or confidential;
5. inter-agency or intra-agency memorandums or letters which would not be available by law to a party other than agency in litigation with the agency;
6. personnel and medical files and similar files the disclosure of which would constitute a clearly unwarranted invasion of personal privacy;
7. records or information compiled for law enforcement purposes, but only to the extent that the production of such law enforcement records or information:
   (a) could reasonably be expected to interfere with enforcement proceedings;
   (b) would deprive a person of a right to a fair trial or an impartial adjudication;
   (c) could reasonably be expected to constitute an unwarranted invasion of a person's privacy;
   (d) could reasonably be expected to disclose the identity of a confidential source, including a State, local or foreign agency or authority or any private institution which furnished information on a confidential basis, and, in the case of a record or information compiled by a criminal law enforcement authority in the course of a criminal investigation or by an agency conducting a lawful national security intelligence investigation, information furnished by a confidential source;
   (e) would disclose techniques and procedures for law enforcement investigations or prosecutions, or would disclose guidelines for law enforcement investigations or prosecutions if such disclosure could reasonably be expected to risk circumvention of the law; or
   (f) could reasonably be expected to endanger the life or physical safety of any individual;
8. contained in or related to examination, operating, or condition reports prepared by, on behalf of, or for the use of an agency responsible for the regulation or supervision of financial institutions; or
9. geological and geophysical information and data, including maps, concerning wells.

Any reasonably segregable portion of a record shall be provided to any person requesting such record after deletion of the portions which are exempt under this subsection. The amount of information deleted shall be indicated on the released portion of the record, unless including that indication would harm an interest protected by the exemption in this subsection under which the deletion is made. If technically feasible, the amount of information deleted shall be indicated at the place in the record where such deletion is made.

Privacy Act (5 USC 552a)

The Privacy Act applies to records maintained by a Federal agency in a system of records in which the primary method for data to be retrieved is by full names, social security numbers, or other identifying particulars.

The Privacy Act states "No agency shall disclose any record which is contained in a system of records by any means of communication to any person, or to another agency, except pursuant to a written request by, or with the prior written consent of, the individual to whom the record pertains." The Act goes on to state the following exceptions, which are discretionary:

Disclosures Permitted by the Privacy Act

The Privacy Act permits disclosure without a subject's consent in certain circumstances:

1. to those officers and employees of the agency which maintains the record who have a need for the record in the performance of their duties;
2. required under section 552 (FOIA) of this title;
3. for a routine use (disclosure of identifiable data outside the Department for a purpose compatible with the purpose for which the data were collected);
4. to the Bureau of the Census for purposes of planning or carrying out a census or survey or related activity;
5. to a recipient who has provided the agency with advance adequate written assurance that the record will be used solely as a statistical research or reporting record, and the record is to be transferred in a form that is not individually identifiable;
6. to the National Archives and Records Administration as a record which has sufficient historical or other value to warrant its continued preservation by the United States Government;
7. to another agency or to an instrumentality of any jurisdiction within or under the control of the United States for a civil or criminal law enforcement activity if the activity is authorized by law, and if the head of the agency or instrumentality has made a written request to the agency which maintains the record specifying the particular portion desired and the law enforcement activity for which the record is sought;
8. to a person pursuant to a showing of compelling circumstances affecting the health or safety of an individual if upon such disclosure notification is transmitted to the last known address of such individual;
9. to either House of Congress, or, to the extent of matter within its jurisdiction, any committee or subcommittee thereof, any joint committee of Congress or subcommittee of any such joint committee;
10. to the Comptroller General, or any of his authorized representatives, in the course of the performance of the duties of the General Accounting Office;
11. pursuant to the order of a court of competent jurisdiction;
12. to a consumer reporting agency in accordance with section 3711 f (Title 31 USC 3711f – Money and Finance: Collection and Compromise).

- The Privacy Act does not protect some records with identifiers:

1. Records of dead persons;

2. Records of individuals who are not U.S. citizens or lawfully admitted aliens;
3. Records not the property of a Federal government agency (Executive Branch);
4. Records not containing full names, social security numbers, or other unique identifiers;
5. Records containing names but not primarily filed and retrieved by name or social security number (e.g., job announcements).

Confidentiality Assurance (Public Health Service Act §308(d), 42 USC §242m(d))

Under the authority in Section 308(d) of the Public Health Service Act, CDC can provide confidentiality protection to a project when necessary to achieve the project's objectives, and when the respondents would not otherwise furnish valid sensitive information without that assurance. 308(d) protects information collected for a project from being used for any purpose other than the purpose for which is was collected unless the person or establishment from which the data were obtained has consented to such use. Confidentiality assurances protect against disclosures under a court order and provide protections that the Privacy Act does not. For example, the Privacy Act only protects individual participants, but confidentiality assurances can also protect institutions. Confidentiality protection granted by the CDC promises participants and institutions that their data will be shared only with those individuals and organizations listed in the consent form and/or the Assurance of Confidentiality Statement for the project. Projects that involve the collection of sensitive information frequently need confidentiality protection. Sensitive information includes (but is not limited to) data collection on sexual behaviors, drug uses, mental health status, or other information that if released could reasonably be damaging to an individual's financial standing, employability, or reputation.

The full text of Section 308(d) of the Public Health Service Act follows:

> "No information, if an establishment or person supplying the information or described in it is identifiable, obtained in the course of activities undertaken or supported under section 304, 306, or 307 may be used for any purpose other than the purpose for which it was supplied unless such establishment or person has consented (as determined under regulations of the Secretary) to its use for such other purpose; and in the case of information obtained in the course of health statistical or epidemiological activities under section 304 or 306 , such information may not be published or released in other form if the particular establishment or person supplying the information or described in it is identifiable unless such establishment or person has consented (as determined under regulations of the Secretary) to its publication or release in other form."

Certificates of Confidentiality - (Public Health Service Act, §301(d), 42 USC §241d)

Under section 301(d) of the Public Health Service Act, CDC can provide a certificate of confidentiality to a research project to protect the privacy of individual participants. Researchers who are authorized to protect the privacy of such individuals may not be compelled in any Federal, State, or local civil, criminal, administrative, legislative, or other proceedings to identify such individuals, without the individual's consent.

The full text of Section 301(d) of the Public Health Service Act follows:

> "The Secretary may authorize persons engaged in biomedical, behavioral, clinical, or other research (including research on mental health, including research on the use and effect of alcohol and other psychoactive drugs) to protect the privacy of individuals who are the subject of such research by withholding from all persons not connected with the conduct of such research the names or other identifying characteristics of such individuals. Persons so authorized to protect the privacy of such individuals may not be compelled in any Federal, State, or local civil, criminal, administrative, legislative, or other proceedings to identify such individuals."

Applying for Confidentiality Protections

Investigator(s) formally apply for confidentiality protections if they believe such protection is necessary to achieve the project objectives and to obtain valid information of a sensitive nature. The detailed instructions and application forms for obtaining **Confidentiality Assurances (308(d)**) and **Certificates of Confidentiality (301(d))** can be obtained from the CDC Management Analysis and Service Office (MASO) website at http://intranet.cdc.gov/maso/confidentiality/confass.htm or by contacting the CDC Confidentiality Officer, at 404-498-1506.

FOIA Request

The CDC FOIA staff, Office of Executive Secretariat, is the focal point for all CDC FOIA requests. The FOIA Officer is the sole official with delegated authority to release or deny CDC records. FOIA requests received by a CDC CIO should be sent immediately to the CDC FOIA Officer, Office of the Executive Secretariat, MS-D54, to be logged and processed. For more information about FOIA contact the CDC FOIA Officer at 404-639-7272.

Table B1: Summary of FOIA, the Privacy Act, 308d, and 301d

Freedom of Information Act	• Obligates Federal agencies to disclose certain information upon a written request • An agency may withhold portion of records under 9 exemptions
Privacy Act	• Protects identifiable records held by Federal agencies from improper disclosures • Protects some identifiable records held by contractors of a Federal agency • Permits disclosures without consent under 12 exemptions
308(d) Confidentiality Assurance	• Rigorously protects individuals' privacy in research and nonresearch projects • Rigorously protects institutions in research and nonresearch projects
301(d) Certificate of Confidentiality	• Rigorously protects individuals' privacy in research projects

The above laws can be found at www4.law.cornell.edu/uscode/.

Table B2: Applicability of FOIA, the Privacy Act, 308d, and 301d

	Contracts	**Grants**	**Cooperative Agreements**
Privacy Act Applies to all records held at CDC in a "system of records."	**Applicable** If the contract calls for a new system of records or extensive additional data that would require establishing a new system. **Not applicable** If the contractor is adding to his/her already established record system.	Not applicable	Not applicable
Freedom of Information Act Applies to all records held at CDC	Applicable if the records are held at CDC		
308d – Confidentiality Assurance	CDC uses for contracts		
301d – Certificate of Confidentiality		CDC uses for grants and cooperative agreements	

12. Supplemental Reading List

Privacy Protection, General

Gostin LO, Hodge JG, Valdiserri RO. Informational privacy and the public's health: The Model State Public Health Privacy Act. *Am J Public Health*. 2001;91(9):1388-92.

Federal Committee on Statistical Methodology, Confidentiality and Data Access Committee. *Confidentiality and Data Access Issues Among Federal Agencies*. Washington DC: Statistical Policy Office, Office of Management and Budget; 2001. Available at www.fcsm.gov/cdac/brochur10.pdf.

O'Brien DG, Yasnoff WA. Privacy, confidentiality and security in information systems of state health agencies. *American Journal of Preventive Medicine* 1999;16(4):351-8.

Sweeney L. Weaving technology and policy together to maintain confidentiality. *Journal of Law, Medicine & Ethics* 1997;25(2&3):98-110.

Horlick GA. Confidentiality. In: Task Force for Child Survival and Development All Kids Count Program, CDC National Immunization Program. *Community Immunization Registries Manual*. Atlanta: CDC; 2000. Available at www.cdc.gov/nip/registry/dl/cirman2.pdf.

Gostin LO, Hodge JG. *Model State Public Health Privacy Act*. Product of Model State Public Health Privacy Project, sponsored by Centers for Disease Control and Prevention, Council of State and Territorial Epidemiologists, Association of State and Territorial Health Officials, and National Conference of State Legislatures. Washington DC: Georgetown University Law Center; 1999. Available at www.critpath.org/msphpa/privacy.htm.

Alpert SA. Health care information: Access, confidentiality, and good practice. In: Goodman KW, editor. *Ethics, Computing, and Medicine: Informatics and the Transformation of Health Care*. Cambridge: Cambridge University Press; 1998. p. 75-101.

National Research Council, Committee on Maintaining Privacy and Security in Health Care Applications of the National Information Infrastructure. *For the Record: Protecting Electronic Health Information*. Washington DC: National Academy Press, 1997. Available at www.nap.edu/books/0309056977/html/index.html.

Gostin LO, Lazzarini Z, Neslund VS, Osterholm MT. The Public Health Infrastructure: A national review of the law on health information privacy. *JAMA* 1996;275(24):1921-7. Full final report from Georgetown/Johns Hopkins Program on Law and Public Health "Legislative Survey of State Confidentiality Laws, with Specific Emphasis on HIV and Immunization" available at epic.org/privacy/medical/cdc_survey.html.

Cox LH. Protecting confidentiality in small population health and environmental statistics. *Statistics in Medicine* 1996;15(17-18):1895-905.

Fienberg SE. Sharing statistical data in the biomedical and health sciences: Ethical, institutional, legal and professional dimensions. *Annual Review of Public Health* 1994;15:1-18.

Office of Technology Assessment. *Protecting Privacy in Computerized Medical Information*. Washington DC: U.S. Congress; 1993. Report No. OTA-TCT-576.

McLaughlin CC. Confidentiality protection in publicly released central cancer registry data. *Journal of Registry Management* 2002;29(3):84-88.

Disclosure Limitation Methods

Armstrong MP, Rushton G, Zimmerman DL. Geographically masking health data to preserve confidentiality. *Statistics in Medicine* 1999;18:497-525.

DeWaal AG, Willenborg LCRJ. Optimal local suppression in microdata. *Journal of Official Statistics* 1998;14(4):421-35. Available at www.jos.nu/Contents/issue.asp?vol=14&no=4.

Fienberg SE, Makov UE, Steele RJ. Disclosure limitation using perturbation and related methods for categorical data. *Journal of Official Statistics* 1998;14(4):485-502. Available at www.jos.nu/Contents/issue.asp?vol=14&no=4.

Hundepool AJ, Willenborg LCRJ. ARGUS: Software for statistical disclosure control. In: Federal Committee on Statistical Methodology. *Record Linkage Techniques – 1997: Proceedings of an international workshop and exposition*. Washington DC: Statistical Policy Office, Office of Management and Budget; 1997:142-149. Available at www.fcsm.gov/working-papers/hundepool.pdf.

Sweeney L. Computational disclosure control for medical microdata: Datafly System. In: Federal Committee on Statistical Methodology. *Record Linkage Techniques – 1997: Proceedings of an international workshop and exposition*. Washington DC: Statistical Policy Office, Office of Management and Budget; 1997:442-453. Available at www.fcsm.gov/working-papers/latanyas.pdf.

Federal Committee on Statistical Methodology, Subcommittee on Disclosure Limitation Methodology. *Statistical Policy Working Paper 22 – Report on Statistical Disclosure Limitation Methodology*. Washington DC: Statistical Policy Office, Office of Management and Budget; 1994. NTIS Document No. PB94-165305. Available at www.fcsm.gov/working-papers/wp22.html.

Disclosure Risk Assessment Methodology

Federal Committee on Statistical Methodology, Interagency Confidentiality and Data Access Group. *Checklist on Disclosure Potential of Proposed Data Releases*. Washington DC: Statistical Policy Office, Office of Management and Budget; 1999. Available at www.fcsm.gov/docs/checklist_799.doc.

Fienberg SE, Makov UE. Confidentiality, uniqueness and disclosure limitation for categorical data. *Journal of Official Statistics* 1998;14(4):385-97. Available at www.jos.nu/Contents/issue.asp?vol=14&no=4.

Samuels SM. A Bayesian species-sampling-inspired approach to the uniques problem in microdata disclosure risk assessment. *Journal of Official Statistics* 1998;14(4):373-83. Available at www.jos.nu/Contents/issue.asp?vol=14&no=4.

Skinner CJ, Holmes. Estimating the re-identification risk per record in microdata. *Journal of Official Statistics* 1998;14(4):361-72. Available at www.jos.nu/Contents/issue.asp?vol=14&no=4.

13. References

[1] Koo D, Wetterhall SF. History and current status of the National Notifiable Diseases Surveillance System. *J Public Health Management Practice*. 1996;2(4):4-10.

[2] Campbell EG, Clarridge BR, Gokhale M, Birenbaum L, Hilgartner S, Holtzman NA, Blumenthal D. Data withholding in academic genetics: Evidence from a national survey. *JAMA*. 2002;287:473-480.

[3] Doyle P. Lane JI, Theeuwes JJ, Zayatz LV. *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*. Amsterdam, the Netherlands: Elsevier Science, 2001.

[4] National Research Council and Social Research Council. *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics*. 1993 Washington D.C.; National Academy Press.

[5] Federal Committee on Statistical Methodology, Confidentiality and Data Access Committee. *Confidentiality and Data Access Issues Among Federal Agencies*; November 2001. Accessible at: www.fcsm.gov/cdac/brochur10.pdf

[6] Federal Committee on Statistical Methodology, Confidentiality and Data Access Committee. Checklist on Disclosure Potential of Proposed Data Releases. July 1999. Accessible at: www.fcsm.gov/committees/cdac/index.html.

[7] DHHS. CDC Confidentiality. February 1984. Available by contacting the CDC Confidentiality and Privacy Officer, in the Office of the Director, Management Analysis and Services Organization (contact information available at: intranet.cdc.gov/maso/home.htm).

[8] DHHS. *NCHS Staff Manual on Confidentiality*, dated 1997, available at inside.nchs.cdc.gov/frames.htm or by contacting the NCHS Confidentiality Officer.

[9] CDC. *CDC/ATSDR Policy on Releasing and Sharing Data.* April 16, 2003. Manual; Guide CDC- 02. Available at: www.cdc.gov/od/foia/policies/sharing.htm.

[10] Doyle P, Lane JI, Theeuwes JJ, Zayatz LV. *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*. Amsterdam, Netherlands: Elsevier; 2001:4.

[11]Federal Committee on Statistical Methodology, Subcommitee on Disclosure Limitation Methodology. *Statistical Working Paper 22 Report on Statistical Disclosure Methodology*, Section B.2. Tables and Microdata (p 3). Washington, DC: Office of Management and Budget, Office of Information and Regulatory Affairs, Statistical Policy Office. May 1994. Available at: www.fcsm.gov/working-papers/wp22.html .

[12] Dunne T. Issues in the establishment and management of secure research sites. In: Doyle P, Lane JI, Theeuwes JJ, Zayatz LV. *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*. Amsterdam, Netherlands: Elsevier; 2001:297-314.

[13] Seastrom MM. Licensing. In: Doyle P, Lane JI, Theeuwes JJ, Zayatz LV. *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*. Amsterdam, Netherlands: Elsevier; 2001:279-296.

[14] CDC. NCHS Policy on Micro-data Dissemination. Available at: www.cdc.gov/nchs/about/policy/policy.htm.

[15] CDC Information Council. *Public Health Information Network Functions and Specifications, version 1.1*. July 25, 2002 (this draft document is not available on the Internet, but is related to document on the CDC Information Council web site titled: *Public Health IT Functions for Emergency Preparedness and Bioterrorism*. Available at: www.cdc.gov/cic/functions-specs/.

[16]DHHS. Guidelines for ensuring the quality of information disseminated to the public by CDC and ATSDRegistry. Available at: www.hhs.gov/infoquality/cdcinfo2.htm.

[17] Federal Committee on Statistical Methodology, Subcommitee on Disclosure Limitation Methodology. *Statistical Working Paper 22 Report on Statistical Disclosure Methodology*. Washington, DC: Office of Management and Budget, Office of Information and Regulatory Affairs, Statistical Policy Office. May 1994. www.fcsm.gov/working-papers/wp22.html .

[18] Federal Committee on Statistical Methodology, Interagency Confidentiality and Data Access Committee. Checklist on Disclosure Potential of Proposed Data Releases; 1999. Washington, DC: Office of Management and Budget, Office of Information and Regulatory Affairs, Statistical Policy Office. Available at: www.fcsm.gov/cdac/.

[19] DHHS, Office of Civil Rights. Health Insurance Portability and Accountability Act of 1996 and Privacy Rule [45 CFR 164.514(a-c)]. Available at: www.hhs.gov/ocr/hipaa/.

[20] Institute of Medicine, Committee on the Role of Institutional Review Boards in Health Services Research Data Privacy Protection, Division of Health Care Services. *Protecting Data Privacy in Health Services Research*. Washington DC:National Academy Press. 2000.

[21] GAO. Record Linkage and Privacy: Issues in Creating New Federal Research and Statistical Information. April 2001. No. GAO-01-126SP.

[22] Federal Committee on Statistical Methodology, Confidentiality and Data Access Committee. *Identifiability in Microdata Files*. Available at: www.fcsm.gov/committees/cdac/.

[23] Land G. Missouri Department of Health and Senior Services—Confidentiality Data Release Rules. Presented at the 2002 CDC Assessment Initiative and National Association for Public Health Statistics and Information Systems. Available at: www.cdc.gov/epo/dphsi/AI/confiles/day1/Land1.ppt.

[24] DHHS. Agency for Healthcare Research and Quality, Healthcare Cost and Utilization Project interactive query website. Available at: www.ahrq.gov/data/hcup/hcupnet.htm.

[25] ORC Macro, Inc. A guide for public health agencies developing, adopting, or purchasing interactive web-based data dissemination systems. (undated document) Contract # 200-96-0598, Task Order No 23. Available at: www.cdc.gov/epo/dphsi/files/WDDSGuideF3.doc.

[26] CDC, Health Information and Surveillance Systems Board Confidentiality Group. *CDC Confidentiality Policies and Procedures Drafted by the HISSB Confidentiality Work Group for Confidentiality Section of J. Reid Draft of HISSB Security, Confidentiality and Data Access Program*.

[27] U.S. Department of Justice, Office of Information and Privacy. *Freedom of Information Act Guide and Privacy Act Overview*, May 2000. Available at: www.usdoj.gov/foia/04_3.html and www.usdoj.gov/foia/foiastat.htm

[28] O'Brien DG, Yasnoff WA. Privacy, confidentiality and security in information systems of state health agencies. *American Journal of Preventive Medicine* 1999;16(4):351-8.

[29] CDC. HIPAA privacy rule and public health. *MMWR* 2003;52(Su[;):1-20. Available at: www.cdc.gov/mmwr/pdf/wk/mmsu5201.pdf.