National Environmental Public Health Tracking Network

Data Re-release Plan

Version 3.

April, 2017

Environmental Health Tracking Branch

Division of Environmental Hazards and Health Effects

National Center for Environmental Health

Centers for Disease Control and Prevention

Table of Contents

Introduction

This document describes the plan for re-release of data submitted to the National Environmental Public Health Tracking Program (referred to as the Tracking Program). These data are scheduled for re-release through the National Public Portal on the National Environmental Public Health Network (referred to as the Tracking Network). State and local Tracking Grantees will submit the data. Since 2006, these grantees have received funding through a cooperative agreement to implement the Tracking Network and to provide core data and measures for the National Portal. Data may also be submitted by non-grantee data stewards with whom, in some cases, a data sharing agreement has been established.

The purpose of this plan is to describe the principles and procedures for the re-release of data on the National Public Portal and to ensure that data are re-released in accordance with applicable federal and state laws. The plan is consistent with the *CDC/ATSDR Policy on Releasing and Sharing Data* and is based on accepted standard operating procedures and data release practices for the agency.

The *CDC/ATSDR Policy on Releasing and Sharing Data* 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 are shared with particular parties with restrictions. The data are shared as soon as it is feasible to do so, taking into consideration 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. The policy recognizes the importance of evaluating data quality and preparing appropriate documentation related to the data (e.g. method of collection, completeness and accuracy, potential limitations on use), including instructions for non-CDC users on the appropriate use of the data, before re-releasing or sharing the data.

This document will be reviewed annually to determine the need for any revisions. If any proposed revisions create a conflict with existing agreements between the Tracking Program and Tracking Grantees or data stewards, Tracking Grantees and data stewards will be notified of the proposed revisions and provided an opportunity to comment before those revisions are finalized.

Background

In January 2001, the Pew Environmental Health Commission called for the creation of a coordinated public health system to prevent disease in the United States. Specifically, the commission saw the need for a system to track and combat environmental health threats. In Fiscal Year 2002, in response to the commission, the U.S. Congress appropriated funding to the CDC to initiate development of the Tracking Program.

The purpose of the Tracking Program is to establish and maintain a nationwide tracking network to obtain integrated health and environmental data and use it to provide information in support of actions that improve the health of communities. The Tracking Network is a Web-based, secure network of standardized electronic health and environmental data. The major functions of the Tracking Network are to

- enable compilation of a core set of nationally consistent health and environmental data and measures;
- discover, describe, exchange, analyze, and manage data;
- make tools available for managing and analyzing the data; and
- provide environmental public health information to the public.

Key benefits of the Tracking Network include the capability to

- provide timely and consistent information for stakeholders;
- provide access to and ability to integrate local, state, and national databases of environmental hazards, environmental exposures, and health effects;
- enable broad analysis across geographic and political boundaries;

- promote systems that are interoperable across jurisdictions through compliance with standards;
- increase environmental public health capacity at the state and local levels;
- provide the ability to enhance and improve data; and
- provide a secure, reliable, and expandable ability to link environmental and health data.

Data and resources available through the Tracking Network will advance efforts to

- identify populations at risk;
- detect trends in the occurrence of environmental hazards, exposures, and diseases;
- generate hypotheses about the relationship between environmental hazards and disease;
- guide intervention and prevention strategies;
- improve the public health basis for policymaking;
- enable the public's right to know about health and the environment; and
- track progress towards achieving a healthier nation and environment.

Nationally Consistent Data and Measures

The Tracking Network includes a core set of nationally consistent data and measures (NCDM) concerning health, exposures, and environmental hazards. Through collaboration with state and local partners and data stewards, these NCDMs have been developed or adopted for the Tracking Network. Health data in the Tracking Network focus on noninfectious health conditions, such as poisoning by carbon monoxide or lead, asthma and other respiratory disease, cancers, and birth defects. Exposure or biomonitoring data include observations of the presence of an environmental agent or its metabolite in persons, such as lead or cotinine in blood and arsenic in urine. Hazard data may include chemical agents (e.g., arsenic), physical agents (e.g., dust particles), and biologic toxins (e.g., harmful algal blooms) that may be found in air, water, soil, food, or other environmental media. Hazard data may be obtained by direct measurement or estimated using mathematical models.

To determine data needs for the Tracking Network, a group of experts, including data stewards, evaluated the data available at the state and national levels to determine their utility for environmental public health tracking (CDC 2008b). Critical elements from these data were used to develop recommendations for nationally consistent data for the Tracking Network.

Through aggregation of records and restriction in the number of available variables, these nationally consistent data balance as much as possible utility and protection of confidentiality. Many of the recommended nationally consistent data are intended as restricted-access datasets (RADS). They are minimally aggregated and, because of small numbers, may contain potentially identifiable information. Such data are not, however, at

the individual level and do not contain personal identifiers such as name, address, date of birth, or social security number.

The nationally consistent data within the Tracking Network are

- aggregated into standardized stratification schema (e.g., counts of events by event code, by location, by time-period, by demographic strata), summarized (e.g., average for location by time-period), and used to calculate derived measures (e.g., age-adjusted rates for location by time-period);
- re-released to the public through the Public Portal as nationally consistent public use data and measures (details are provided in related sections of this document);
-
- linked to other health outcomes data, exposure and biomonitoring data, environmental hazards and environmental monitoring data, and other population and census data by association through spatial proximity, temporal proximity, or membership in a population subgroup for the purpose of statistical analysis; and
- analyzed for the purposes of identifying populations at risk, detecting and tracking temporal or spatial trends, generating hypotheses about the relationship between environmental hazards, exposures, and disease, and guiding intervention and prevention strategies and policy development.

Data Transport, Management, and Storage

Data Transport

Data will be transported from the data steward or Tracking Grantee to the CDC Tracking Program in accordance with acceptable practices ensuring the protection, confidentiality, and integrity of the data contents. The Tracking Program supports the following transport mechanisms for receiving data:

- Secure File Upload via CDC Secure Access Management System (SAMS)

SAMS provides a simple, secure way for the transporting of data to the Tracking Network. It is a Web application that does not require any software installation but does require pre-authentication. If a data steward is unable to transport data via PHINMS or SDN SFU, other arrangements will be made for secure data transportation.

Data Format

The Tracking Program preferred format for data received is XML validated against XML schema developed for the particular dataset and made available to the data steward or Tracking Grantee (CDC 2008c).

Data Repository

Data provided to the Tracking Program by data stewards or by Tracking Grantees will be archived, stored, protected, or disposed of in accordance with relevant federal records requirements and applicable information systems management requirements. However, it is the intent of the Tracking Program to store and make data continuously available as described in this plan. Data will be stored in the National Data Repository on CDC servers for secure and sensitive data, with access limited to authorized CDC employees and contractors. Data sharing agreements with data stewards or cooperative agreements with Tracking Grantees may require renewal as specified by the terms of each agreement. Should an agreement be terminated, the Tracking Program will delete all copies of the original data, derived data, and standardized or aggregated data within the Tracking Network, pursuant to the applicable federal records requirements.

Release of Data via the Public Portal

The Public Portal serves as the Tracking Network's interface to the public and will enable stakeholders to access nationally consistent public use data and measures as well as other services and tools offered by the Tracking Network. A nationally consistent measure is a specific combination, calculation, or derivation of health or environmental data that yields a composite number, such as a count or rate for a specific geographic unit and time period of analysis. For purposes of the Tracking Program, public use data sets (PUDS) for health and biomonitoring data are defined as data sets comprised of aggregated data with all individually identifiable data or information removed, and with the remaining fields modified or suppressed to reduce disclosure risk as much as reasonably possible and to prevent the creation of any table that violates the numerator or denominator cell size rules. PUDS for environmental data are defined as data sets comprised of aggregated or summarized data that have been processed to allow for interpretation of the original data and presentation of nationally consistent environmental measures.

The Public Portal will provide users with the capability to

- Access PUDS and measures compiled as nationally consistent data and measures.
    - View prepared reports.
    - View preconfigured tables, charts, and maps containing count data, measures, and smoothed measures.
    - Execute flexible, user-defined queries where health and biomonitoring data are restricted to smoothed or age-adjusted measures.
    - Access the PUDS repository via an open Application Programming Interface (API)
- Search and view metadata.
- Browse and view other relevant Tracking Network data sources, which may include U.S. EPA data, U.S. Census Bureau data, and other CDC data.
- Browse and traverse relevant categorized links to other information sources, including
    - State and local Tracking Programs.

  - State environmental and health agencies.
  - U.S. Environmental Protection Agency.
  - Other CDC Websites.
- Access public health messages as well as descriptive content designed to assist the public in interpreting the NCDMs and other data and measures.

Before data is released on the Public Portal, grantees and other data providers will have the opportunity to use a secure, non-public section of the Tracking Portal to validate their respective data prior to public release.

Confidentiality Protection and Statistical Stability

PUDS and nationally consistent measures for health and biomonitoring data require the protection of confidentiality and notice of statistical stability. PUDS and nationally consistent measures for environmental data do not have the same requirements regarding confidentiality, except with respect to protecting sensitive data related to national security and explaining data limitations and uses.

PUDS and nationally consistent measures for health and biomonitoring data will not contain information that is identifiable or potentially identifiable according to currently accepted procedures for reducing disclosure risk. To address issues of confidentiality protection and statistical stability, a combination of disclosure control procedures including additional aggregation, suppression, and smoothing will be employed to generate and display PUDS or nationally consistent measures. These procedures will be used to prevent disclosure of non-zero counts less than 6 for geographic units with populations under 100,000 and corresponding crude rates. Rates and other measures will be flagged as unstable when aggregation or smoothing or both fail to produce estimates with a relative standard error (RSE) less than 30%.

Creating PUDS and Measures: Aggregation and Variable Restriction

RADS will be used to generate PUDS or measures for re-release through the Public Portal. This includes the aggregation of data into standardized stratification schema and the calculation of summary or derived measures such as age-adjusted rates for public release. In generating PUDS or measures, several procedures will be used to block breaches of confidentiality and prevent disclosure of confidential information. These procedures include predefined aggregation and variable restriction. Aggregation may be performed spatially or temporally. To reduce the number of small counts, the number of demographic variables provided will depend on the level of aggregation. Data highly aggregated spatially or temporally will contain more demographic variables than data minimally aggregated. The generation of PUDS and measures may also involve the collapsing of variables, (e.g., the collapsing of 5-year age groups into one group for children and one group for adults). The variable or variables chosen for aggregation, restriction, or collapsing will depend on the intended purpose of the PUDS or measure.

Display of PUDS and Measures: Primary and Complementary Suppression

Even after aggregation and variable restriction, when displayed as tables, charts, or maps, PUDS and measures may still contain small case counts, thereby requiring additional steps to protect confidentiality and ensure statistical stability. The primary suppression rule is to censor case counts and corresponding crude rates for geographic units with total populations under 100,000 and positive case counts fewer than 6. Corresponding crude rates will be censored to prevent back-calculation of counts.

After the appropriate suppression rules are applied, all PUDS or measures available as prepared reports or through preconfigured data products will be evaluated for necessary complementary suppression to prevent back-calculation of initially censored counts. When complementary suppression is necessary, the following hierarchy will be followed:

1. First, attempt to find a candidate cell representing a small (population < 100,000) geographic unit;
2. Next, seek a candidate cell among a large geographic unit; and
3. Lastly, suppress the relevant marginal total.

The goal underlying this approach is to preserve marginal totals (e.g., statewide totals) whenever possible— they are of the greatest general interest and likely to be available via other Web sites or publications.

Although data products may be individually secured through primary and complementary suppression, it may be possible to reconstruct censored data by combining information from multiple data products. In a flexible query system allowing multiple dynamic queries, implementing effective suppression rules for sets of dependent data products is often difficult. To ensure protection of suppressed counts, sets of preconfigured, dependent data products will be evaluated collectively to determine whether additional confidentiality protection procedures are needed. When sets of dependent data products are too large or difficult to ensure protection of suppressed counts, health and biomonitoring data will be presented as smoothed crude measures, smoothed age-specific measures, or age-adjusted measures.

Once the suppression process has been completed, some rates may remain that are statistically unstable. Instability can arise from small numerators (case counts) or small denominators (populations or subpopulations). Any rate or measure with a RSE ≥ 30%, will be flagged as unstable, but not suppressed.

Smoothing

When appropriate, smoothing will also be used to protect confidentiality, increase statistical stability, and provide information in areas where counts or unsmoothed measures may otherwise be suppressed or flagged. Smoothing combines data from multiple geographic units and can yield greater stability in local rate estimates. Smoothing and flagging unstable measures can be combined to address situations in

which even the smoothed estimates are not considered sufficiently stable. Appropriate public health messages and interpretation guidance will accompany any presentation of smoothed measures. Such messages will clarify that smoothed measures are intended to present general trends rather than accurate local measures.

For re-release of PUDS and nationally consistent measures to the public, smoothing approaches will

- be relatively easy to implement using available computing software;
- execute quickly so that they can be employed in conjunction with dynamic queries; and
- use calculations which although algebraically straightforward, are difficult to reverse.

As an example, a simple empirical Bayes (EB) smoothing procedure (Waller and Gotway 2004) may be suitable. Such an approach combines geographic data in a manner that is sensitive to rate variations across geographic units by giving greater weight to immediately local estimates as the relative variability in the surrounding rates increases.

Disclosure Statement

The Public Portal will prominently display a public release disclosure statement informing users of their responsibility to maintain confidentiality and inform CDC immediately if, during use of data obtained from the Public Portal, the identity of an individual person is inadvertently disclosed. In addition, it will state that users will not imply or state that interpretations based on the data are those of the original data source or CDC but will acknowledge both in all reports based on these data.

Metadata and Instructions to Users

Data released through the Public Portal will be accompanied by the necessary documentation in the form of metadata about collection procedures, completeness, and limitations. Metadata describes the content, quality, and context of the data and provides links to additional information such as quality assurance documents and data dictionaries. Each dataset will have a corresponding metadata record created by or with the assistance of the data provider. Metadata will be maintained in a central repository and revised as data are updated.

The creation and maintenance of metadata is a vital component of the Tracking Network. They will be used to generate instructions on the proper use and interpretation of data available on the Tracking Network. Metadata also support additional Network functionality by providing users the ability to

- locate and access data based on key fields within the metadata such as title, purpose, abstract, keywords, geographic boundaries, year, and content, and
- discover data and evaluate its quality, limitations, restrictions, and appropriateness for the intended use.

Public Health Messages

Data released on the Public Portal will be accompanied by several types of messages. The portal will include contextual information that explains how the content area and its data and measures are related to environmental public health tracking. This includes a discussion of current understanding about the association between health and the environment. Information about data limitations and appropriate uses of the data and measures will be provided as will guidance on interpretation. Applicable national level measures and relevant national objectives will be provided, such as those in Healthy People. Where appropriate, the Public Portal will also contain public health messages related to prevention and environmental health stewardship. Public health messages will be developed in collaboration with subject matter experts from multiple state, federal, and nongovernmental partners. Public health messages developed for the Public Portal will be consistent with CDC's existing public health messages.

Asset Protection

CDC understands that many of the recommended, nationally consistent data are intended as restricted-access datasets in that they are minimally aggregated and may contain potentially identifiable information. CDC also understands that it may receive other information from data sources that may be deemed proprietary and/or commercial, or data that may be subject to protections from disclosure provided by specific privileges.

Consistent with applicable federal laws, regulations, and policies, CDC intends to use its best efforts and the procedures set out in this Data Re-Release Plan to

- protect the privacy and confidentiality of any potentially identifiable information,
- protect any proprietary or commercial information provided to it by a Tracking Grantee, national source or data source other than state or local Tracking Programs, and
- protect any other data exempted from disclosure under the Freedom of Information Act (5 U.S.C. Sec. 552) or other applicable federal laws or privileges.

CDC also intends that if requested or required to disclose information outside of the procedures provided for in this Data Re-Release Plan, CDC will provide the respective data source or sources with prompt written notice of any such request or requirement, thus allowing the data source or sources to assert any applicable privilege or position related to the disclosure of the information.

Glossary

Term	Definition
Authentication	The process by which the identity of a person requesting access to restricted data or services is verified.
Authorization	The process by which a person's right to access restricted data or services is verified.
Confidentiality	The treatment of information entrusted to CDC with the expectation that it will not be divulged to others in ways that are inconsistent with the conditions agreed to when the information was originally disclosed.
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.
Data re-release	Re-release of data provided to CDC
Data sharing agreement (DSA)	A mechanism by which a data requestor and CDC can define the terms of data access that can be granted to requestors.
Data Steward	The person(s) responsible for the management, processing, documentation, integrity, and security of information in a data system.
Disclosure	Unauthorized public disclosure of information about a person, about which data have been collected.
Disclosure control	Procedures used to limit the risk that information about an individual or potentially identifiable information will be disclosed. These procedures include restricting access, aggregation, variable restriction, suppression, and smoothing.
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.
Metadata	Data about data that describes the content, quality, and context of the data and provides links to additional information like quality assurance documents and data dictionaries.
Nationally Consistent Data	Specific data collected, organized, and in some cases pre-processed, on the basis of standards adopted by CDC for the Tracking Network
Nationally Consistent Measure	A specific combination/ calculation/ derivation of health and/or environmental data to yield a composite number, such as a rate for

	a specific geographical unit and time period of analysis. Standard measures have been developed/adopted for the Tracking Network through collaboration with state and local partners and data steward.
National Data Repository	A repository of Tracking Network data that will be used to store, process, and make available data and measures
Tracking Network	A web-based, secure, distributed network of standardized electronic health and environmental data
National Public Portal	A portal (Website with special features) that will be accessible to the public to access data and other resources on the Tracking Network
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.
CDC Secure Access Management System (SAMS)	A CDC electronic authentication service that provides application access for external partners and the secure exchange of electronic files between CDC and partner organizations.
Public-use datasets (for environmental data)	Data comprised of aggregated and/or summarized data which have been processed for the purposes of interpretation and presentation of nationally consistent environmental measures.
Public-use datasets (for health and biomonitoring data)	Data that are comprised of aggregated data 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.
Restricted-access datasets	Data that (1) are minimally aggregated for increased utility, (2) do not contain personal identifiers, (3) contain potentially individually identifiable data as a result of small numbers, and (4) are shared only with authorized users.
Smoothing (spatial)	A statistical approach in which a local measure (e.g., a rate) is adjusted by referring to measures from surrounding geographic units. A key objective is to produce stable estimates.
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.
XML	A standard coding language that allows information and services to be encoded with meaningful structure and semantics that computers and humans can understand. XML supports information exchange and can be extended to include user-specified and industry-specified tags
XML schema	The structure for constraining the contents of XML documents

Bibliography

Centers for Disease Control and Prevention. CDC/ATSDR Policy on Releasing and Sharing Data. Atlanta, GA: Centers for Disease Control and Prevention; 2005. Available at https://www.cdc.gov/maso/policy/releasingdata.pdf.

CDC Funding Opportunity Announcement CDC-RFA-EH17-1702: Enhancing Innovation and Capabilities of the Environmental Health Tracking Network. April 2017. Available at: https://www.cdc.gov/nceh/tracking/pdfs/cdc-rfa-eh17-1702.pdf.

CDC-CSTE Intergovernmental Data Release Guidelines Working Group. CDC-ATSDR Data Release Guidelines and Procedures for Re-release of State-Provided Data. Atlanta, GA: Centers for Disease Control and Prevention; 2005. Available at: https://stacks.cdc.gov/view/cdc/7563/.

CDC. Environmental Public Health Tracking Network: Data Transport How-to-Guide. Atlanta, GA: Centers for Disease Control and Prevention; 2008a. Available at: <TBD>

CDC. Environmental Public Health Tracking Network: Recommendations for Nationally Consistent Data and Measures. Atlanta, GA: Centers for Disease Control and Prevention; 2008b. Available at: <TBD>

CDC. Environmental Public Health Tracking Network: Schema Documentation. Atlanta, GA: Centers for Disease Control and Prevention; 2008c. Available at: <TBD>

CDC. National Environmental Public Health Tracking Program. Technical Network Implementation Plan. Atlanta, GA: Centers for Disease Control and Prevention; 2007. Available at http://www.cdc.gov/nceh/tracking/pdfs/TNIP_V1.pdf.

User Guide for CDC's SAMS Partner Portal. Atlanta, GA: Centers for Disease Control and Prevention; 2017. Available at https://auth.cdc.gov/sams/SAMSUserGuide.pdf?disp=true

Federal Committee on Statistical Methodology. Statistical Policy Working Paper 22 (Second Version, 2005): Report on Statistical Disclosure Limitation Methodology. Available at: http://www.fcsm.gov/committees/cdac/.

Pew Environmental Health Commission. 2000. America's Environmental Health Gap: Why the Country Needs a Nationwide Health Tracking Network: Technical Report. Baltimore, MD: Pew Charitable Trusts. Available at: http://healthyamericans.org/reports/files/healthgap.pdf

Waller LA, Gotway CA. *Applied Spatial Statistics for Public Health Data*, Chapter 4. New York: Wiley. 2004:86-98.