# CSTE Data Release Reference Document

Tabulated, Aggregated, and Macro Data

CSTE

**Policy Development Recommendations for Public Health Agencies**

**Developed by the CSTE Data Release, Presentation, Access, Suppression, and Sharing (DRPASS) Workgroup**

# Acknowledgments:

Abigail Cheney, MPH, Michigan Department of Community Health;

Taylor Jones Pinsent, MPH, Council of State and Territorial Epidemiologists;

MaryKate Martelon, MPH, Massachusetts Department of Public Health;

Laura Hayes, PhD, Connecticut Department of Public Health;

Suzanne N. Gibbons-Burgener, DVM, PhD, Wisconsin Department of Health Services;

Ian Painter, PhD, Washington State Department of Health;

Eman Addish, MPH;

Louise Saw, MPH, CIC Central Utah Health Department; and

Ashley Bush, DrPH

*Inclusion in the acknowledgments does not imply endorsement of all content or viewpoints expressed in this document, and the views presented do not necessarily reflect the official policies or positions of all workgroup members, their affiliated institutions, or any organizations acknowledged herein.*

# Contents

# Executive Summary

Developed by the Council of State and Territorial Epidemiologists (CSTE) Data Release, Presentation, Access, Suppression, and Sharing (DRPASS) Workgroup, this policy development reference document is designed to assist public health agencies in creating or updating data release policies for public health data in aggregate formats. Aggregation of data reduces granularity by summarizing data points and minimizes the risk of identifying specific individuals or small groups while still enabling valuable insights.

This document was developed in response to the growing need among applied public health practitioners, attorneys, researchers, media, elected officials, and community-based wellness providers for clear, actionable guidance on sharing or accessing public health data with sufficient granularity (e.g., demographic data). Such data are critical for effectively communicating with policymakers, securing funding, and addressing health inequities. However, workgroup discussions confirmed anecdotal reports that many agencies lack current data release policies or continue to rely on outdated frameworks. We hope this document will support the development of clear guidance on how to support this need while maintaining the individual privacy of the people represented within the data.

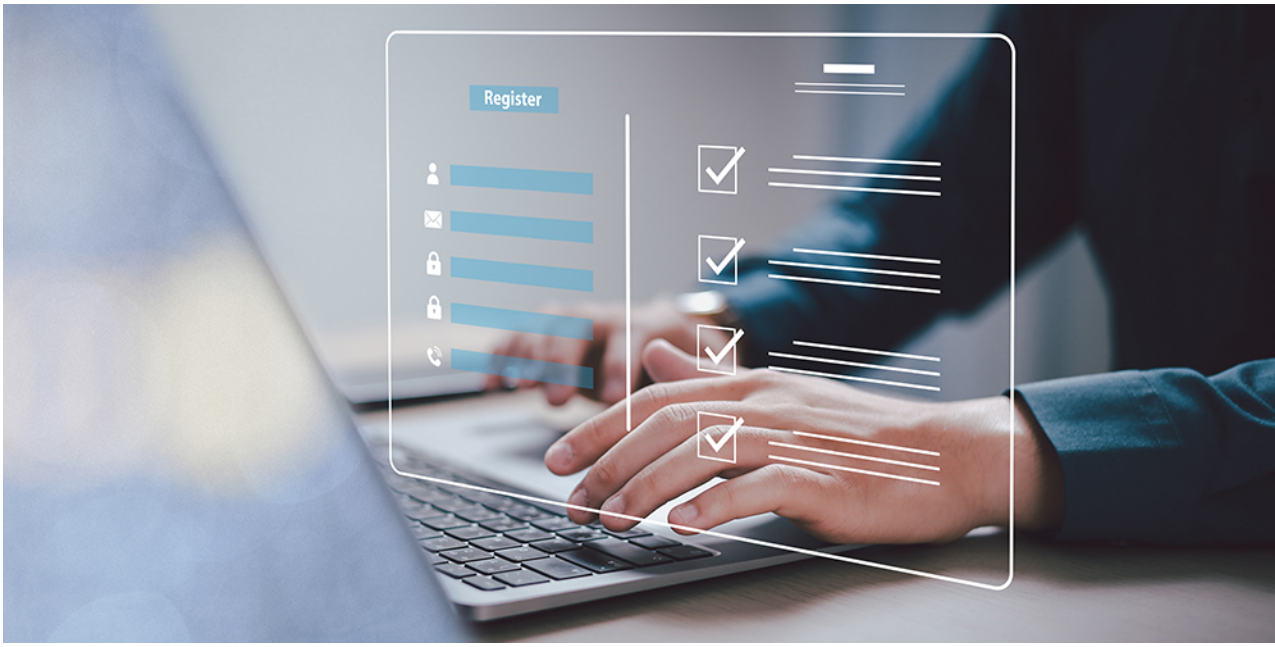
The document takes a holistic approach, providing considerations and recommendations across the **data release lifecycle**, integrating **data equity and legal checkpoints** at each stage. These checkpoints offer non-legal staff practical guidance for developing equitable and legally sound policies. By embedding complex legal and ethical concepts into an accessible format, this document empowers public health agencies to balance data utility with privacy and equity considerations.

This document is poised to enhance public health action and awareness, helping agencies share data responsibly, with technical accuracy and methodological rigor in data interpretation, while addressing health disparities and supporting community data needs. It positions data release policies as pivotal in advancing public health goals through informed decision-making and equitable resource allocation. More importantly, this document is not intended to add restrictions or create data access barriers.

# Purpose and Scope of Document

This reference document was developed by the Council of State and Territorial Epidemiologists (CSTE) Data Release, Presentation, Access, Suppression, and Sharing (DRPASS) workgroup. It contains considerations, examples, and suggestions for public health agencies that want to develop guidelines and policies for releasing public health surveillance data in an aggregate format. Data release often occurs in response to custom data requests or routine release of aggregate data through formal reporting mechanisms (e.g., dashboards, static reports). Please note, this document includes general recommendations **without** regard to the type of data (e.g., vital statistics, infectious disease, hospital discharge), the format of data (e.g., stand-alone datasets, static dashboard, dynamically updated dashboard), or the requestor of the aggregate data (e.g., external partner or media, intra-agency, programmatic). It is meant for applied epidemiologists, data stewards, or anyone who manages or advises on the use or release of public health data.

This document does not provide recommendations for the release of record-level data or for data submitted to CDC for national situational awareness or re-released by CDC. It is not intended to be the national standard for aggregate data release. The authors of this document recognize the variation of release practices across the country. However, the workgroup does hope some elements of these recommendations, as well as the definitions outlined within, are useful as a guide for jurisdictions to update their own policies. It should be viewed in service of an agency's broader data governance efforts.



# Background

The Council of State and Territorial Epidemiologists (CSTE) is a professional association of applied epidemiologists at state, Tribal, local, and territorial (STLT) public health agencies. CSTE is organized under six main steering committees, with corresponding subcommittees that focus on specific topics.

Since late 2018, the CSTE Surveillance Policy Subcommittee has discussed considerations related to data suppression and release of aggregate data. During those discussions, the subcommittee voiced the value of having CSTE develop an aggregate data release reference document that health departments could use to develop their own state and local policies. In March 2018, the CSTE Data Release Workgroup was created to develop a guidance and reference document for public health agencies to use for their own policies for aggregate data. During a workshop at the 2019 CSTE annual conference, the group highlighted several gaps in the current draft guidance. As a result, an assessment was sent out to state epidemiologists to understand the landscape of release and suppression policies. Before fully integrating the assessment results into the guidance and reference documents, the workgroup was put on hold in 2020 to respond to the global COVID-19 pandemic.

In late 2023, a new workgroup with an expanded scope was formed: the CSTE Data Release, Presentation, Access, Suppression, and Sharing (referred to as DRPASS or as the workgroup moving forward) workgroup. Based on the previous workgroup's experience and the current workgroup's feedback, it was decided to continue the previous charge of developing a reference document that could be used as a guide for jurisdictions creating or updating their own release policies. This workgroup, however, decided to take a holistic approach and aims to spotlight considerations and recommendations at each stage of an aggregate data release lifecycle for public health action and awareness.

# Methods

The CSTE DRPASS workgroup meetings, along with the Surveillance Policy Subcommittee (SPS) meetings, served as the primary source of information and guidance for this document. These meetings reached over 1,000 current CSTE members working in public health agencies across the country. The DRPASS workgroup aimed to address various aspects of the data release lifecycle through comprehensive discussions at each stage. The workgroup employed multiple formats such as speaker sessions, expert panels, and breakout rooms for in-depth conversations over 12 months. From these discussions and notes, considerations and recommendations were developed by a writing group.

For the data privacy section, external consultants (i.e., Privacy Hub) and experts were engaged to lead a series of focus groups, with the goal of gathering insights on what content would be most valuable for developing effective training materials related to needs, challenges, and priorities of different health department staff involved in data collection, management, release, and privacy. The privacy focused sessions included a mix of participants from various sectors, such as public health departments, academia, and industry (e.g., data analytics companies, private organizations involved in data handling and privacy solutions), to ensure a wide range of perspectives and expertise.

# Document Definitions

For this document's purposes, recognizing that organizations may employ different definitions, please review the following definitions of commonly used terms. The following terms were identified by workgroup members and defined using a range of sources.

The term **"data"** as used throughout this document, refers to collections of qualitative or quantitative information. In the context of public health, data are used to describe features, health characteristics and status, or other information. This information could be at the level of event, individual, establishment, population, or associated environment. Data are broadly defined to include vital records and statistics, reportable conditions and events, health survey information, laboratory-derived information, social determinants of health, personally identifiable information (PII), and non-PII.

### Aggregate data

sometimes referred to as tabular data, are generated from record-level data by grouping the data according to specific attributes of records (e.g., demographic or geographic attributes) and creating summary measures for these groups, often in tabular reports. Common summary measures include the count of records in each group, the count and proportion of records which meet specific criteria, and other descriptive statistics, such as the mean or median value of a specific attribute.

### Anonymization

is the process that removes the association between the identifying dataset and the data subject. This data privacy technique irreversibly removes all identifying information.[1]

### Data democratization

is the act of making information accessible to people, including assuring that they have adequate tools and resources to understand and make use of the data

### Data equity

is a set of principles and practices to guide the collection, analysis, interpretation, and access to data to support health equity.

### Data release policy

is a set of rules and guidelines on how and when to release data outside an organization.

### Data utility

refers to the usefulness of data for its intended purpose, including how well it supports analysis, decision-making, or other applications.

### De-identification

as defined under HIPAA (the Health Insurance Portability and Accountability Act of 1996), is the process by which identifiers are removed from health data to mitigate privacy risks to individuals. This process helps to support secondary use of data (e.g., for comparative effectiveness studies, policy assessments, and life sciences research). HIPAA-compliant de-identification of protected health information (PHI) is possible using two methods: Safe harbor and expert determination.

1 https://nvlpubs.nist.gov/nistpubs/ir/2015/nist.ir.8053.pdf

**De-identified data**

do not include any direct identifiers, and the indirect identifiers included (e.g., dates, demographic information) are sufficiently generalized such that the risk of re-identifying individuals included in the dataset is low. In practice this means that (1) the data do not contain direct personal identifiers, including but not limited to name, address, telephone number, Social Security number (SSN), health identification number, or other identification numbers, and (2) protected health information cannot be determined for any individual through the use of indirect identifiers, such as combinations of individual characteristics or demographic information, either within a dataset or in combination with other de-identified data.

**Direct identifiers**

are data elements that can provide a direct link to an individual's identity, such as elements directly associated with an individual's identity (e.g., name, SSN, phone number, address, biometric identifiers, digital identifiers).

**Health equity**

is the state in which everyone has a fair and just opportunity to obtain their highest standard of health. Where all individuals have equitable access to resources, opportunities, and healthcare necessary to achieve their full health potential, irrespective of their socioeconomic, racial, or geographical backgrounds.[2]

**Indirect/quasi-identifiers**

are data elements which are not direct identifiers but can be combined with other information to reveal a person's identity or protected health information. These include elements such as event dates, hospitalization dates, geographic region, and demographic information.

**Metadata**

are data about data and refers to the information required to process, analyze, interpret, or otherwise make meaningful use of a dataset. It may include technical details about the structure and format of data elements (such as those found in a data dictionary), as well as contextual information. Metadata serves as the "blueprint" or "context" for understanding and appropriately using data.

**Open/public data**

is defined as data available to the public or data which meet an appropriate risk threshold to release to the public. Open/public data may be released through an open data portal (e.g., data.cdc.gov), as part of a dashboard, or through other means. Generally open/public data are expected at a minimum to be de-identified.

**Personally identifiable information (PII)**

refers to data that can directly or indirectly be used to identify, locate, or contact individuals or can reveal the characteristics or other details about them. Such data are regulated by general data privacy laws (e.g., the Family Educational Rights and Privacy Act (FERPA)). Examples include name, phone number, date of birth, and home address.

**Protected health information (PHI)**

is a subset of PII that specifically relates to a person's health information created or used by a covered entity (e.g., hospital, doctor, or insurer) and regulated under HIPAA. This includes medical records, billing information, lab results, and treatment history.

2 Prentice KR, Beitelshees M, Hill A, Jones CH. Defining health equity: A modern US perspective. iScience. 2024 Nov 5;27(12):111326. doi: 10.1016/j.isci.2024.111326. PMID: 39640575; PMCID: PMC11617406.

**Pseudonymization**

is a form of de-identification that replaces identifying information with pseudonyms, breaking direct links to individuals while allowing authorized public health agency analysts to re-identify data when necessary using securely stored keys.

**Public health surveillance data**

are any data collected or used for the purpose of monitoring the health of a population, including health, provision of care, and risk factors.

**Re-identification**

identification of data involves leveraging de-identified data elements, in combination with publicly available information, in order to discover the person to whom the data relates.

**Restricted data**

are data that may have some restriction or sensitivity (e.g., identifiable or partially identifiable public health data) but can be shared with a use/sharing agreement and approval.

**Small number and data suppression guidelines**

are rules and practices used by programs or agencies to limit the release of data to protect the privacy of individuals represented in the data. These guidelines help determine when data should be withheld or modified, especially when releasing small counts could lead to the identification of individuals. Guidelines include information on reliability recommendations, disclosure risk, cell suppression rules, and small-number restriction recommendations and standards by data type (e.g., population data, survey data).

**Statistical disclosure control (SDC)**

also known as statistical disclosure limitation or disclosure avoidance, is a technique used to ensure that no person or organization is identifiable from the results of an analysis of survey or administrative data or the release of microdata.[3]

**Unavailable data**

are data not made available to the public either as open/public data or through a use/sharing agreement or other process. This may be due to policy or legal restrictions, insufficient data quality, lack of digital format (e.g., paper copies only), or other concerns.



3 https://en.wikipedia.org/wiki/Statistical_disclosure_control

# Data Release Lifecycle

The data release lifecycle (Figure 1) outlined in this document formalizes data release steps for public health practitioners in various capacities, including data stewards and analytics teams. This process begins with the intake, review, and evaluation of aggregate data requests or report development. Once these preliminary steps are completed, the cycle progresses to data privacy considerations and the preparation of the interpretation or narrative to be released alongside the data. The lifecycle concludes with final agency approval and dissemination, often aligning with STLT mission statements or strategic planning activities.

This lifecycle aligns with CSTE's 2024–2026 Strategic Impact Plan, specifically Priority #2: Advance health for all by engaging communities in data collection, access, analysis, and use. It also falls within the framework of the 10 Essential Public Health Services (EPHS),[4] originally established in 1994 and updated in 2020. In particular, the lifecycle supports EPHS #3: Communicate effectively to inform and educate people about health, the factors that influence it, and ways to improve it, as well as EPHS #9: Improve and innovate public health functions through ongoing evaluation, research, and continuous quality improvement.

Considerations and recommendations are offered during each stage of the lifecycle, with references to **data equity** and **legal checkpoints** within each stage. A checkpoint offers considerations for the reader to reflect on and may offer a course of action depending on the answer. These were developed and are displayed throughout the lifecycle to add an equity lens and address legal considerations. This document is meant to offer guidance at any point during the lifecycle, so please reference the stage that suits the specific need. For simplicity, this document offers recommendations intended for use within specific programs, but these should not be viewed as limited to the program level—they may also be applicable more broadly across agencies or systems.

### Figure 1. Data release lifecycle


4 https://www.cdc.gov/public-health-gateway/php/about/index.html

# Considerations for Data Release Policy Development

This section walks through each stage of the data release lifecycle. By integrating considerations at each stage, agencies can build processes and a workforce that is well-equipped to manage aggregate data requests, mitigate re-identification risks, and maintain a balance between privacy, equity, and data utility. This approach will promote more effective and equitable data practices that align with both legal requirements and community expectations.

## Data Equity Checkpoints

Many state, Tribal, local, and territorial (STLT) public health agencies demonstrate their commitment to data equity through mission statements, health assessments and improvement plans. They may also have established practices for providing data that support health promotion and disease prevention. At the same time, health agencies typically have clear mandates to prioritize health programming and initiatives that are impactful to populations most vulnerable to poor health outcomes in their jurisdictions.

**Data equity** conceptually intersects these overarching strategic objectives. An emergent framework[5] within the public health field, data equity is a set of principles and practices to guide the collection, analysis, and interpretation of data throughout the data lifecycle to support **health equity**, the state in which everyone has a fair and just opportunity to obtain their highest standard of health.

To the extent that STLT health agencies have identified and prioritized data equity, additional considerations, such as those put forward by the CDC Foundation[6] specific to data access, can be relevant to the data release lifecycle (the scope of this document).

**Examples of those considerations include the following:**

- Provide resources, processes, and avenues to facilitate dispositioning of data in accordance with participants' wishes. (Nelson et al, 2020)
- Publicize clear data release schedules and information on where to go and how to access data once released. (Nelson et al, 2020)
- Adhere to data management best practices for data access, including clear data destruction parameters following use. (Nelson et al, 2020)
- Carefully de-identify and anonymize data prior to release.
- Use role-based permissions for data access to ensure security and confidentiality. (Nelson et al, 2020)
- Develop an accessible data request process with clear policies and procedures for evaluating requests.
- Clearly document if/why data are unavailable (e.g., legislation, data quality, preparation) (Nelson et al, 2020)
- Include community partners in defining which data should be released.

5 https://weallcount.com/the-data-process/

6 The bulleted text is drawn from p. 15 of "Principles for Using Public Health Data to Drive Equity," published by the CDC Foundation. https://www.cdcfoundation.org/data-equity-principles?inline

Additionally, **data democratization** is a foundational principle within the broader context of the data release lifecycle[7]. It emphasizes making data accessible and understandable to all individuals—regardless of their background in analytics, IT, or software—so they can engage meaningfully with the information. In line with data democratization, requests for aggregate data release may present opportunities to build health literacy and inclusion for people who are underserved by state or local governments. An example is prioritizing data requests from communities historically or currently overburdened with health inequities when programmatic resources for timely fulfillment of all requests are limited. Decisions regarding the extent of interpretation that accompanies filled data requests may also be informed by the data equity lens.

## Compliance Obligations and Legal Checkpoints

The legal landscape governing the release of aggregate public health data varies significantly across health departments due to a complex interplay of federal regulations, state statutes, Tribal sovereignty, and local ordinances. While aggregate data are generally considered less sensitive than individual-level data, their release remains subject to varying legal and jurisdictional standards.

At the federal level, frameworks such as the Health Insurance Portability and Accountability Act (HIPAA) establish baseline privacy protections but do not preempt more stringent state or Tribal laws. States and territories may impose additional legal requirements or policies that govern the de-identification, use, and public dissemination of health data, including thresholds for minimum cell sizes to prevent re-identification, specific authorization processes, or data-sharing agreements. Local health departments often operate under the authority of state law but may adopt policies reflecting local priorities and community sensitivities.

Tribal health authorities, as sovereign entities, maintain independent legal systems and data governance protocols rooted in Tribal law, cultural values, and sovereign rights, including the protection of community-specific health information.

As a result, compliance obligations for the release of aggregate public health data exist along a spectrum—from open data access policies to highly restrictive policies requiring rigorous legal and ethical review. Understanding and navigating this spectrum is essential to ensure lawful, respectful, and equitable data sharing practices that promote public health goals while protecting community trust and individual privacy.

7 drawn from "Principles for Using Public Health Data to Drive Equity," published by the CDC Foundation. https://www.cdcfoundation.org

# Stage 1: Intake

The first stage in the data release lifecycle pertains to data request intake and processing. Covered topics include the use of a standardized tool (data request form) for intake of data requests, classification of requestor information (requestor affiliation), request tracking, and target time frames for fulfillment of data requests.

## Data Request Form

Health agencies receive requests for aggregate data in various ways, ranging from simple verbal inquiries about data availability to formal written requests. An aggregate data request form (DRF) is considered a key asset for a health agency for the collection of necessary information to fulfill a data request and can mitigate the need for excessive follow-up communications with requestors. Additionally, elements captured on a DRF can be used to track request types, sources, and volumes, as well as to monitor compliance with programmatic data delivery requirements across an agency.

DRFs can be made available through various means, such as via an online portal, web-based application, fillable electronic or printed form, or spreadsheet. The CSTE DRPASS workgroup has compiled a library of example DRFs that can be accessed through the CSTE Connect platform. CSTE Connect requires a user to establish a free account and profile to gain access to user/member-only resources and discussions on the platform. Consequently, hyperlinks will require the user to be logged in to view those documents housed in CSTE Connect which can be found here.

**Common elements to include in a DRF are:**

- Requestor information: name, affiliation, affiliation type, contact information
- Description of proposed data use or data request purpose (as applicable)
- Asking requestor to indicate if data will be used for research (i.e., determine the need for Institutional Review Board (IRB) review)
- Public health database/dataset or program from which data will be extracted to fill the request (Consider including a link to available data.)
- Requested data time frame (year(s)) and geographic area (e.g., state, county, region)
- Indicator for resident records only or for all occurrences within geographic area (e.g., all residents of a state versus all events occurring within the state regardless of the person's residence)
- Variables requested, including units and subcategories as selected from a list of available variables
- For tables: by variable and levels
- For tables: counts, rates, percentages, or a combination
- Acknowledgment of processing fees (as applicable, including requests for multiple jurisdictions)
- Requested time frame for data product delivery (justification)
- Requested file format (e.g., sas7bdat, .csv, .xlsx) and mode of data product transfer (e.g., SFTP, email/secure mail)
- Names of individuals who will have access to the data and to whom the data visualizations or analyses will be released
- Signature (as applicable)

To support data transparency and accessibility of publicly available datasets, public health agencies can post DRFs, along with completion and submission instructions that include range options for variables and years available, on publicly accessible websites. In many cases, requestors may need assistance with form completion and submission. Details on the appropriate contact for such support should accompany the form. Additionally, guidance on appropriate use and completion of a DRF should be provided. **Considerations including the following:**

- **Use guidance** educates requestors about the types of data that can be formally requested using the form, as well as redirection guidance for submission of other types of requests not within scope of the DRF. Basic details on the distinction between aggregate and individual person or incident data types may be informative.
- **Completion guidance** informs requestors on correct completion of each of the elements of the form; it should be provided either directly on the form or as part of the web-based DRF posting. For tabular data product requests, guidance on how to designate the desired levels or subcategories of a variable in the requested table is helpful.

Completed DRFs should be directed to assigned staff or a centralized intake email repository that staff monitor. Staff may be assigned to conduct an initial form review to confirm its completeness and any preliminary review tasks necessary to verify that information provided is sufficient for initial request routing. An auto-response that, minimally, acknowledges receipt of the DRF and provides an expected time frame for follow-up communication is recommended.

### Requestor Affiliation

Multiple decision points in the data request lifecycle, such as the expected request fulfillment time frame and the extent of provided data interpretation, often depend on the data requestor's affiliation. Fields in DRFs can be customized to collect affiliation information from requestors to enable simple categorization. Classification of requestor affiliations (see examples below) can inform processing of requests throughout the data request lifecycle. **Examples include:**

- General/public
- Media
- Federal agency or Tribal nation
- Intra-agency (internal to public health agency or legislators)
- Inter-agency
- Inter-jurisdictional (different jurisdictional authorities)
- Academic/research organization
- Community-based organization
- Other for-profit/NGO/501(c)(3)

Most often, requestor affiliation will be based on self-reporting, as provided in a DRF. To the extent that the affiliation type of the requestor influences subsequent processing steps for a given request, verification of affiliation may be undertaken as an additional initial review step. Additional subclassification after intake may be useful to the agency's processes.

## Request Tracking

Another fundamental component to the intake step in the data request lifecycle is the initiation of a tracking record. Formalized tracking of aggregate data requests received by a public health program, from the point of receipt of the initial request to final close-out, supports application of Plan-Do-Study-Act (PDSA),[8] a recommended framework of quality improvement within the data release lifecycle. Further, tracking data requests align with the emerging use of business intelligence (BI) tools in public health agencies.

Multiple data request tracking tools are available, including spreadsheet applications (e.g., Excel, Access), basic open-source software products, and more advanced data governance software applications (e.g., GovQA, REDCap, MS Power Platform tools).

**Candidate core elements to include in a data request fulfillment record include:**

- Unique number linking DRF and fulfillment record
- Requestor's contact information
- Modifications to initial request and date of DRF completion
- Date of DRF receipt
- Decision to fill, redirect, or deny (with the reason) a request and date of decision
- Date final data product was delivered to requestor
- Storage location of agency copy of the final product and documentation of methods used to create final product (e.g., data source, code, analyses, references)

Additionally, links to more detailed files pertaining to each request in the tracking system may eliminate redundancies in fulfilling repeat data requests or similar requests. Documentation of internal decision approvals at various stages and steps of the data release lifecycle can be incorporated into a tracking system. During intake of a DRF, it can be useful to document the decision to fill, redirect, or deny a request. More robust tracking systems used for PDSA may collect additional elements such as intended data use and delivery method and can be used to support analytics of requestors' data-asset use.

## Time Frame to Fill Requests

Data requests submitted to public health agencies are often time sensitive. Establishment and clear communication to requestors of data-product delivery time frames support customer-service oriented business operations of public health agencies. Communication to requestors of an expected time frame for data delivery may be more straightforward than establishment of time frames themselves.

**Considerations for establishing an estimated data request delivery date include:**

- Agency standards (and staff capacity) for timeliness of responses in internal and external communications
- Responsiveness of requestors to follow-up communications
- Statistical programming and analytical time associated with request preparation
- Buffer time needed for final data product review and approval
- Application of tools to provide required confidentiality protections, including a secure method of transfer and file encryption
- Time to prepare narrative interpretation of data products, as appropriate

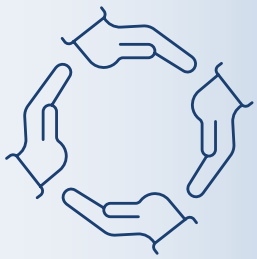
8 https://www.ncbi.nlm.nih.gov/books/NBK599556/

Total time needed to accurately fill data requests may vary widely due to these factors. **Best practices for timely data request fulfillment are as follows:**

- Respond to communications from data requestors to clarify details of the request and request DRF resubmission within one or two business days.
- Confirm decision as to whether to fill requests and the expected delivery date within two or three business days of finalizing DRF details.
- Deliver data products within the agreed upon time frames.
- Communicate to requestors within one or two business days any expected delays on a final decision to fill a request or on the data product delivery.

## Data Equity Checkpoints

- Create a standardized intake tool.
- Ensure tool can be read by screen reader (accessibility of form). Consider other accessibility, language and 508 compliance.[9]
- Consider including a method to request assistance for filling out form.
- Consider including a method to request assistance for analysis and interpretation.
- Ensure that requestor classification categories are inclusive of the variety of people and groups that request data.[10]
- Use the request tracking system to analyze and monitor themes related to affiliation partners and what types of data are being requested. Use the request tracking system to analyze timeliness by type.







9 https://www.section508.gov

10 Classification can smooth access. For example local or tribal governments may have rights to unsuppressed data, so that, if the data request is going to a state, data suppression might not be needed.

# 2 Stage 2: Review and Evaluate

The second stage of the data release lifecycle focuses on reviewing requests for custom data products and assessing the resources needed to fulfill them. The goal is to decide whether the request moves on to Stage 3: Data Privacy. This review often involves creating preliminary tabulations, consulting requestors or agency leadership for clarification, and exploring alternative options when data release limitations arise. It's essential to have clear agency protocols for how the data requests are reviewed and who is consulted for certain types of requests. Coordination with legal staff may also be required, as detailed in the "legal checkpoints." Due to these factors, this stage can be the most resource intensive.

## Taking Inventory

A logical first step in reviewing a data request is to consider the availability of the requested data products. It's not unusual for requestors to ask for data products that have already been released by the program. A simple first question is, "Are these data already available?" Programs may find themselves contacting other programs and state agencies to address this first question. Tools from federal agencies, for example CDC Wonder,[11] often serve as a satisfactory redirection for data requestors. The availability of a searchable inventory of datasets and databases within an agency can facilitate determining if data are already available to the public, whether relevant information is collected by the agency, and to whom to direct questions (such as the data steward) for data that the agency collects that are relevant to the request. Access to searchable tools provides transparency to the community as well as helps reduce the effort required on the part of the data steward or request handler. When requested data are not published elsewhere, the follow-up question becomes, "Can we produce the requested data products?" Three themes—data utility, availability, and infrastructure—may factor into this final determination.



11 https://wonder.cdc.gov

2

## Data Utility

Public health agencies seek to release data that will serve the purpose of their intended use. **Dimensions of data utility to consider during review and evaluation include:**

- **Fitness of purpose:** Are the requested data suitable for addressing the stated purpose of the request?
- **Quality:** Do data fields, on which tabulations will be based, meet standards of quality that are set by either the health agencies themselves or federal agencies in terms of completeness and accuracy for the intended purpose? If such standards are not documented, is the steward who is filling the request confident that data quality concerns will not negatively influence interpretation of aggregate data products?
- **Timeliness/concurrency:** Are the datasets that will be used current within the request time frame? If not, is an earlier time frame sufficient to meet the requestor's needs, or can a complete, supplemented data product be provided at a later date?
- **Small numbers:** Will data products include cell values that cannot be released due to re-identification risk associated with small numbers? If so, will the data products that remove sensitive cell values still serve the original intended purpose of the request? For more information on how to address these questions see the data privacy stage.
- **Interpretation:** To what extent can (or should) interpretation of aggregate data be included as part of request fulfillment?

In some situations, small-number considerations may result in aggregate data that are suppressed to the extent that the data are rendered not useful for the intended user. In this situation, consider directing the requestor to the process for requesting limited-use data (which might still be aggregate data) that requires a data use agreement.

Additionally, achieving an appropriate balance between minimizing risk of disclosing confidential information and preserving the utility of the data for its intended use is an ongoing challenge in statistical disclosure control (SDC). STLT agencies must adopt context-sensitive SDC approaches that are legally compliant, ethically sound, and tailored to the specific analytic needs of data users. This involves careful trade-offs between privacy and precision and calls for transparent policies, partner engagement, and a clear articulation of the data's intended purpose.

## Data Availability

Data availability must be assessed and articulated to the requestor. It is possible that data requested are already publicly available on state websites or other agency sites (e.g., CDC Wonder) and a redirection to those resources is sufficient. Alternatively, the data requested might not be collected, either in part or in their entirety. If that is the case, then a conversation or redirection may be warranted. A conversation might be between a data steward and the requestor to amend the request based on the guidance of agency staff. A redirection might send the requestor to another agency that may be able to provide the information.

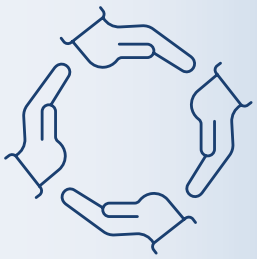
## Data Infrastructure

State and local public health agencies are often faced with resource and staff limitations. Additionally, requests that require extensive analytics for informed analysis and interpretation may exceed the expertise available within a program area. In such cases, decisions whether to fill requests at all or to fill such requests outside of response-time standards may be necessary. Transparent and documented protocols that allow for denial of requests due to resource shortages and capacity limitations may be reasonably developed and approved within health agencies.

2

## Data Equity Checkpoints

- Data utility
  - If quality or small numbers are an issue, consider including a checkpoint to reach back out to the requestor. For example, discuss with the requestor the privacy risk and recommendations for combining cells or other mechanisms to obtain data, such as data use agreements (DUAs) or data sharing agreements (DSAs).
  - When including data interpretations with a data request, consider if additional plain language checks should be conducted.
- Infrastructure
  - Consider the benefits and risks of creating structures to cover the expenses associated with fulfilling data requests. For example, should a fee waiver process be available? What validation process might need to be in place to confirm requestors are with waiver-eligible agencies?
  - Investment in the development and maintenance of searchable inventories creates a more equitable and transparent data request process.



## Legal Considerations Checkpoints

- Aggregate data generally carry minimal re-identification risk, except when small numerators or denominators are disclosed.
- **The following are legal and compliance mechanisms that can support additional release of relevant data:**
  - Does release of aggregate data require Institutional Review Board (IRB) review? IRB is not often required for aggregate data release, unless the request includes merging multiple datasets and in so doing enables the identification of individuals whose data are analyzed. Also consider if an IRB was required to collect the information considered for release and, if so, what stipulations does that IRB put in place for further data release?
  - Are formal agreements (e.g., memoranda of understanding, data use agreements) used in the collection of the data, and do they place limitations on data requests and release?
  - More formalized agreements may be needed in the case of the release of small numbers for limited use. This could include, for example, the release of aggregate data to Tribal Health Authorities that necessitates the inclusion of small numbers.
  - Intra- and inter-agency governance rules may apply and may require development of formal agreements between entities sharing and receiving aggregate data. If formal agreements are needed, then guidance from this stage may not apply.

# Stage 3: Data Privacy

This stage in the data release lifecycle centers on data privacy implications and considerations when developing a data privacy policy.

Two major recommendations for building and supporting data privacy in public health jurisdictions were raised during focus groups (see Appendix A). These recommendations included (1) development of standard operating procedures (SOPs), guidance, or policies for data management and (2) workforce training.

## Recommendation #1: Guidance, Standard Operating Procedures (SOPs), or Policies for Data Management

The first—and essential—recommendation, when developing guidance, SOP, or policies for data management, particularly concerning data privacy and release, is to create frameworks that are comprehensive, clear, and adaptable to various contexts.

**Here are some factors to consider during the data privacy policy development process.**

### 1 Consideration for Who Gets a Voice in Designing the Policy

- **Inclusive policy design:** Ensure that the policy development process includes a diverse range of partners. This can include representatives from public health, academia, and community groups, as well as data privacy experts. This inclusivity helps ensure the policy reflects multiple perspectives and addresses the needs and concerns of all partners.
- **Partner engagement strategies:** Use surveys, focus groups, workshops, or public comment periods to gather input from different groups. Make sure that people who are underserved by state or local governments have meaningful opportunities to contribute.

### 2 Consistently Interpretable and Easy to Understand

- **Plain language:** Use clear, non-technical language to explain data privacy concepts. Avoid jargon and provide simple explanations for technical terms.
- **Visual aids:** Incorporate flowcharts, diagrams, and visual aids to illustrate processes and decision points, making complex ideas more accessible to a broad audience.
- **Tiered explanations:** Develop materials with varying levels of detail to cater to different audiences. For example, provide a simple overview for the public and a more detailed explanation for data professionals or academics.

## 3 Allow for Replicable Results

- **Standardized procedures:** Create clear, standardized procedures for data collection, analysis, and reporting to ensure that results are consistent and replicable. This can include specifying the methods for data suppression and aggregation.
- **Documentation of methods:** Clearly document all data-handling processes, including how decisions about data suppression or modification are made. Provide examples of how these procedures have been applied in practice.
- **Transparency:** Share data processing guidelines and methods publicly so partners understand how results are derived and can replicate the analysis if needed.

## 4 Account for Varying Sensitivity of Data Being Reported

- **Data sensitivity classification:** Develop a framework for categorizing data based on its sensitivity (e.g., highly sensitive, moderately sensitive, non-sensitive). This will guide decisions on the level of protection required for different types of data.
- **Tailored data release strategies:** Define specific strategies for each sensitivity level, such as more stringent suppression rules for highly sensitive data (e.g., personal health information) and less stringent rules for aggregated or anonymized data.
- **Dynamic policies:** Ensure that policies are adaptable to changing contexts and can be updated to reflect new data-sensitivity considerations, such as emerging risks or changes in data privacy laws.

## 5 Account for Varying Levels of Statistical Sophistication Required

- **Customized guidance:** Develop guidance documents tailored to different levels of statistical expertise. Provide basic instructions for those with limited statistical knowledge and more advanced guidelines for professionals.
- **Training and support:** Offer training sessions, webinars, and support materials to help all audiences understand data privacy and statistical concepts. Consider creating a help desk or consultation service to assist with complex data requests.
- **Clear communication:** Ensure that all communications are accessible to those with varying levels of statistical understanding. Use examples and case studies to illustrate key points.

## 6 Consider the Conditions Under Which Data Are Reported

- **Context-specific policies:** Take into account the political, social, and cultural context of the data release.
- **Scenario planning:** Develop scenarios that outline different conditions (e.g., public health emergency declaration) and specify how data release policies should be adjusted accordingly.
- **Community affected awareness:** Make sure the community affected are aware of the context in which data are being reported, including any limitations or potential risks associated with the release.

3

### 7 Ensure Policies Are Understandable to the Target Audience

- **Audience-specific messaging:** Tailor messages to different audiences. For public health partners, emphasize the relevance to health outcomes. For community members, focus on transparency and trust. For the media, provide clear, concise summaries with supporting data.
- **Feedback mechanisms:** Use surveys, focus groups, or public consultations to test whether policies are understood by different audiences and adjust based on their feedback.
- **Accessible formats:** Ensure that policy documents are available in multiple formats (e.g., print, digital, audio) and languages to reach diverse audiences.

## Balancing the information and privacy trade-off

Appropriately balancing information/privacy trade-off is an essential focus within a data management policy because poorly implemented de-identification can potentially lead to bad decisions and bad science (for example, complete privacy protection hinders data reporting, and high-resolution quasi-identifiers within a dataset lead to poor privacy protection).

To assist epidemiologists and data reporters in balancing data release (with privacy protection, several key aspects need to be navigated. These critical aspects include essential concepts in data privacy, statistical disclosure control, regulatory considerations, communication methodologies, and future developments (e.g., artificial intelligence). These were synthesized and developed from workgroup breakout discussions.

**Here's an overview of these critical aspects that need to be navigated:**

### 1 Essential Concepts and Key Elements in Data Privacy, Security Risk, Assessing Risk, and Reporting

- **Data privacy versus data equity:** Understand the trade-offs between making data accessible (data equity) and protecting individual privacy. Data equity aims to ensure fair access to data, whereas privacy focuses on protecting personal information.
- **Re-identification:** Use methods like concatenation or grouping or those specifically designed for data visualizations to minimize the risk of re-identifying individuals from published data.
- **Identifiability:** Quantify how easy it is to re-identify individuals from the data.
- **Sensitivity:** Measure the potential harm or invasion of privacy that could occur if data are disclosed, including personal or medical information.
- **Safeguards:** Put technical and procedural protections in place to secure the data and mitigate the risks of breach or re-identification.
- **Privacy risk controls:** Set limits and conditions on the uses and disclosures that may be made of such information without an individual's authorization.
- **Security risk controls:** Implement appropriate administrative, physical, and technical safeguards.
- **Data minimization:** To reduce the potential for privacy breaches and assist with de-identification, disclose only the data necessary for a specific purpose and for compliance with data agreements and memoranda of understanding.

- **Limited data set (LDS):** An LDS is a HIPAA-specific term and refers to health information that has had all direct identifiers (e.g., names, addresses, phone numbers, SSNs) removed, but may still contain certain potentially identifying elements such as dates and some geographic information. An LDS may be disclosed to an outside party without the individual's authorization only when:
  - The disclosure is solely for research, public health, or health care operations; and
  - The recipient is typically required to enter into a Data Use Agreement (DUA), Data Sharing Agreement (DSA), or Data License Agreement (DLA) with the disclosing entity. Note: Although HIPAA allows disclosure of an LDS under these conditions, some organizations or state laws may impose stricter requirements, such as Institutional Review Board (IRB) approval, explicit patient authorization, or other agency-specific protocols.
- **Data use agreement (DUA):** A DUA, sometimes referred to as a Data Sharing Agreement (DSA), is a type of legal agreement used for sharing data, and may be required by a disclosing entity before sharing data with a recipient.. It establishes clear parameters for permissible uses of data, outlines the responsibilities of data users, and specifies safeguards required by most institutions. A Data License Agreement (DLA) is a specific type of DUA where a data provider grants a user/organization limited rights to a dataset, retaining ownership but governs how data can be accessed and used for a specific purpose.

## 2 Key Concepts of Statistical Disclosure Control (SDC)

- **Application of SDC:** SDC is a set of methods and techniques used to protect the privacy of individuals when releasing statistical data (see Appendix B). The goal of SDC is to minimize the risk of disclosing sensitive information about individuals, households, or organizations while preserving the usefulness of the data for analysis and decision-making. Some techniques to protect confidential information within released datasets include suppression/restriction, aggregation, and data swapping . Apply these techniques to epidemiological data to prevent identification of individuals while maintaining data utility.
- **Disclosure risk:** Disclosure occurs when an individual's or entity's identity or attributes are inadvertently revealed through released data.
- **Types of disclosure:**
  - **Identity disclosure:** When the released data allow someone to determine the identity of an individual or entity in the dataset (e.g., through unique combinations of attributes)
  - **Attribute disclosure:** When sensitive information about an individual or entity is revealed, even if their identity is not directly disclosed (e.g., revealing health status or income)
- **Data utility:** Assess the degree to which the data remain useful and accurate for analysis after applying SDC techniques.
- **Risk assessment:** Regularly assess the risk of re-identification, especially in datasets with small sample sizes or unique data combinations.
- **Key elements to consider when assessing risk:**
  - Identifiability
  - Sensitivity
  - Safeguards
  - Resources

## 3 Regulatory Considerations

- Understand whether your organization is subject to certain privacy laws including:
  - Privacy Act of 1974:[12] This law governs the collection, maintenance, use, and dissemination of PHI that is maintained in systems of records by federal agencies.
  - Health Insurance Portability and Accountability Act (HIPAA):[13] This law governs the use and protection of personal data by public and private entities. HIPAA sets a national baseline for health data privacy, but jurisdictions can adopt stricter privacy laws (e.g., the California Consumer Privacy Act). To determine whether your organization is subject to HIPAA or other privacy laws, you should consult with your internal legal/privacy team. There are resources[14] available to help public health practitioners and their respective attorneys understand how HIPAA applies and the steps that must be taken to become a HIPAA hybrid entity.
- **HIPAA Safe Harbor method (including drawbacks of the method):** Safe Harbor is a specific section of the HIPAA Privacy Rule that specifies the removal of 18 types of variables within a dataset. Remediation in this way requires no expert knowledge and is straightforward to implement but has some limitations. The exact requirements for Safe Harbor, as per 45 CFR § 164.514, can be found on the HHS website.[15]

  **Some drawbacks to this method of de-identification include:**
  - **Cannot always cope with the complexity of modern datasets:** The rules for Safe Harbor were written in the mid-1990s and struggle to cope with the size and scope of modern datasets.
  - **Does not remediate all potential "risky" variables.**
  - **Utility loss:** Blanket removal of elements based on broad categorization can result in removal of data that add little to no risk but which could have a lot of utility (e.g., removal of all dates more specific than the year, which are data that could be very useful in determining time between events: diagnosis, recurrent symptoms, infectious disease incubation periods, etc.). Additionally, restrictions on elements like geographical granularity can hinder use cases (e.g., epidemiologic analysis at a county or census tract level).
  - **No consideration of dataset as a whole:** While the variables identified in this method can confer risk to a dataset, if only a limited number of these variables exist in combination, the overall risk of re-identification can be relatively low. Safe Harbor does not weigh the risk-versus-utility trade-off here, as noted above.
  - **No independent assessment:** No external expert assessment is performed as part of the Safe Habor method, which can be problematic when it comes to auditing procedures, as there is no independent documentation that processes are compliant.

12 https://www.justice.gov/opcl/privacy-act-1974
13 https://www.hhs.gov/hipaa/for-professionals/privacy/index.html#:~:text=The%20HIPAA%20Privacy%20Rule%20establishes,care%20providers%20that%20conduct%20certain
14 https://www.networkforphl.org/resources/an-overview-on-conducting-a-hipaa-hybrid-entity-assessment-for-local-public-health-departments/
15 https://www.hhs.gov/hipaa/for-professionals/special-topics/de-identification/index.html#:~:text=The%20second%20is%20the%20"Safe%20Harbor"%20method:&text=(ii)%20The%20covered%20entity%20does,strategies%20that%20minimize%20such%20loss

3

- **HIPAA expert determination method (including drawbacks of the method):** This method requires a person with appropriate knowledge of and experience with generally accepted statistical and scientific principles and methods of rendering information that is not individually identifiable to assess the dataset in question. The expert is required to document in the methods and results of the analysis that there is a very small risk that the information could be used, alone or in combination with other reasonably available information, by an anticipated recipient to identify an individual.

  **Some drawbacks to this method of de-identification are as follows:**

  - **Who constitutes an "expert" is not specifically defined:** The HIPAA Privacy Rule does not provide robust definitions for the qualifications required for an individual to be deemed an "expert," and the exact methodology used to assess the risk associated with a dataset can vary between experts. Experts can be drawn from fields such as statistics, mathematics, and science and are required to have professional experience or academic or other training in the use of de-identification methodologies.
  - **Guidance is less clear-cut:** Determination of the acceptable level of risk is open to interpretation, as the HIPAA Privacy Rule specifies only that datasets must have a "very small" level of risk.
  - **Assessment validity must be maintained:** Expert assessments take into consideration factors such as changes to technology, social conditions, and the availability of information over time. While this can be of benefit to long-term utility, it also means that many experts apply a validity date to their assessments, after which point the data will require reassessment.

- **Factors that contribute to the risk associated with a dataset:** These include the concepts of "anticipated recipients," "reasonable availability" of external information, and "very small risk" under HIPAA. (Note: Mathematically, insistence on zero re-identification risks sets an unachievable goal.)
- **HIPAA de-identification provisions:** Even when the HIPAA Privacy Rule does not apply, the HIPAA de-identification provisions are often used to help provide guidance regarding statistical disclosure risk analyses and associated confidentiality. The privacy principles that can be drawn from HIPAA de-identification provisions can help jurisdictions balance the mandate to protect the public's health with the other important goal of maintaining public trust in the agency's attentiveness to individuals' confidentiality.
- **Hybrid entity:** A public health authority (PHA) may designate itself as a hybrid entity under HIPAA if it performs both covered and non-covered functions. In this case, only the "health care component" (i.e., the part of the agency performing HIPAA-covered activities) is subject to HIPAA requirements. Other parts of the agency remain HIPAA-exempt but may still receive data from covered components within the hybrid entity. This designation allows internal data sharing from HIPAA-covered components to non-covered components within the same agency, facilitating public health activities while limiting the scope of HIPAA compliance. However, the re-release or further disclosure of data by the non-covered components is governed by other applicable federal, state, or local laws, not HIPAA. HIPAA's de-identification provisions are often used to help meet these broader legal and regulatory requirements and to support public trust in responsible data practices.

- Family Educational Rights and Privacy Act (FERPA):[16] This is a federal law that affords parents the right to have access to their children's education records, the right to seek to have the records amended, and the right to have some control over the disclosure of personally identifiable information from the education records. When a student turns 18 years old, or enters a postsecondary institution at any age, the rights under FERPA transfer from the parents to the student ("eligible student"). The FERPA statute is found at 20 U.S.C. § 1232g, and the FERPA regulations are found at 34 CFR Part 99.
- **Other state and federal agency privacy rules:** Public health and environmental data collected or managed by state and federal agencies are subject to a range of privacy and confidentiality rules designed to protect individuals, communities, and sensitive locations. These rules vary by agency, data type, and legal authority, and they often extend beyond personal identifiers to include protections for sensitive geographic, agricultural, or environmental information.

### 4 Evaluation and Communication Methodologies

- **Strengths versus weaknesses:** Compare strengths and weaknesses of different approaches to the quantification of risk in a dataset (to learn more, see the CSTE Data Privacy Training which include examples of k-anonymity, sample uniques (SU), population uniques (PU), and the probability of being a population unique given sample uniqueness (e.g., Prob PU|SU)).
- **Transparent communication:** Clearly communicate data limitations and the steps taken to protect privacy. This builds public trust and ensures informed use of the data.
- **Data interpretation:** Provide context for data findings, highlighting potential biases or limitations due to privacy measures.
- **Technological advances:** Stay up to date on emerging technologies, such as differential privacy and artificial intelligence. Related technologies are transforming how data are collected, interpreted, and used—but also amplify privacy concerns. The intersection of AI and data privacy demands thoughtful governance, ethical design, and a combination of technical safeguards and legal protections to ensure responsible innovation.
- **Policy changes:** Monitor evolving regulations and guidelines, which may impact data sharing and privacy protocols.

## Recommendation #2: Workforce Training Topics and Considerations

The second recommendation raised by the focus groups is workforce training. To address this, CSTE developed a Data Privacy Training. The training is designed to help epidemiologists navigate essential privacy concepts, statistical disclosure skills, and relevant regulatory considerations. In addition to this training resource, jurisdiction should develop a data privacy focused workforce training program that **include the following topics:**

### 1 Principles of Re-identification Risk

- **Understanding re-identification**: Explain how seemingly anonymous data can be re-identified by combining it with other data sources. Use real-world examples to illustrate how re-identification can occur and the potential consequences.
- **Factors affecting re-identification risk:** Discuss the key factors that increase the risk of re-identification, such as data granularity, sample size, and the availability of external data.

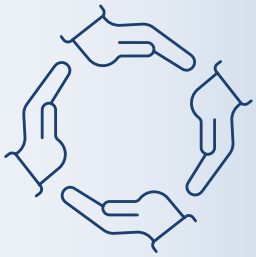
16 https://www.cdc.gov/phlp/php/resources/family-educational-rights-and-privacy-act-ferpa.html

## 2 Geography and Time in Relation to Re-identification Risk

- **Spatial and temporal risks:** Provide training on how geographic detail (e.g., ZIP codes, census tracts) and time frames (e.g., daily versus annual data) can increase re-identification risk, especially in small populations or unique events.
- **Mitigation strategies:** Teach methods to mitigate these risks, such as geographic aggregation, temporal smoothing, or using broader time frames.

## 3 Data Equity Checkpoints

- Discuss how privacy practices can impact different groups, particularly for people who are historically underserved by state or local governments. Provide scenarios where privacy needs might conflict with equity goals and discuss strategies for balancing these concerns.
- Train staff on ethical considerations when deciding how to balance privacy with the equitable use of data. Provide frameworks or checklists to guide decision-making in complex situations. **For example:**
  - Be mindful of the word "suppression" and what it means to your audience. Are you hearing a negative connotation associated with the word? Try using "withheld due to maintaining confidentiality/privacy" instead.
  - Can you explain the reason for suppression in a non-technical way? Clearly state why data have been suppressed or restricted, such as small sample sizes, privacy concerns, or potential for re-identification. For example, if some subpopulations were aggregated, include a note: "Data for certain subpopulations are aggregated to protect individual privacy due to low numbers of individuals included."
  - Acknowledge how suppression may affect the visibility of certain groups. Use language that emphasizes the importance of the people in a particular subpopulation even though their specific data are not presented. For instance, "While data are not displayed for people within certain groups due to small sample sizes, we recognize the importance of the people within these subpopulations and continue to monitor and address their specific needs."
  - Include a section in reports or dashboards that details the suppression policies or criteria (e.g., counts of fewer than five individuals are suppressed). Explain how these policies are intended to protect privacy but may also impact the visibility of small numbers. Can you link to any policies?
  - Discuss how suppression aligns or conflicts with equity goals. For example, "To protect individual privacy, data for people within certain groups have been suppressed. This may limit our understanding of the full scope of disparities affecting these people, and we are exploring ways to address this gap while maintaining privacy."
  - In any table, graph, or report where data are suppressed, add detailed footnotes explaining why specific data points are missing. Example: "Data for people within certain subpopulations have been suppressed due to low counts and confidentiality concerns. This decision is in line with our data protection policy but may under-represent the true prevalence within these groups."
- Consider using disclaimers at the beginning or end of the data presentation that explicitly acknowledge that suppression may disproportionately impact the visibility of people who are under-served by state or local governments.



## Legal Considerations Checkpoints

### Understanding Legal Rules and Constraints Related to Privacy and Re-identification Risk

- **Privacy risk laws and regulations:** Provide an overview of relevant laws and regulations. Cite the legislation associated with the quantification of risk in patient-related datasets, namely HIPAA Privacy and Security Rules and Centers for Medicare and Medicaid Services' (CMS) cell size rules.
- **Differentiate between privacy and security risks:** Privacy risks relate to appropriate safeguards to protect PHI and set limits and conditions on the uses and disclosures that may be made of such information without authorization from the individual. Security risks relate to the implementation of appropriate administrative, physical, and technical safeguards.
- **Compliance requirements:** Train staff on how to ensure data practices comply with relevant laws, including understanding exemptions, data-sharing agreements, and mandatory reporting requirements. Also, ensure data policies have appropriate administrative safeguards to ensure auditing of access.

### Data Suppression Options and Best Practices

- **Data suppression techniques:** Teach various data suppression techniques, including small-cell suppression. Employ case studies to demonstrate how each method can be applied in different contexts.
- **Develop standards for data suppression:** Collaborate with partners to create standards for data suppression tailored to the agency's context. Ensure these standards are clear, consistent, and well documented.
- **Implement best practices:** Build out best practices for data suppression that balance privacy protection with data utility. Include guidelines on how to determine when suppression is necessary, how to document suppression decisions, and how to communicate these practices to data users.

As mentioned earlier, to support the privacy training needs of jurisdictions, CSTE developed a Privacy Training. The training comprises three courses (with 11 lessons total), and it is designed to help navigate essential privacy concepts, statistical disclosure skills, and relevant regulatory considerations that are required for optimally balancing epidemiological public health practice data reporting responsibilities with privacy/confidentiality protection. This training is free and can be accessed on CSTE Learn.[17] The audience is anyone who creates or reports epidemiological data findings to the wider community.

The CSTE Privacy Training was created in response to recommendations made in the Network for Public Health Law Disaggregation of Public Health Data by Race & Ethnicity: A Legal Handbook.[18] To facilitate data disaggregation, specifically for race and ethnicity data, the handbook highlights the need to train public health authority staff in statistical disclosure risk assessment and control. It further recommends recruiting or training individuals on the Safe Harbor method as well as accepted statistical and scientific principles and methods for rendering information not individually identifiable.

For those reporting epidemiological data, it's crucial to stay informed about the latest privacy-preserving technologies (i.e., privacy-preserving record linkage) and regulatory changes that shape the landscape of data sharing and protection. Additionally, policies need to be regularly reviewed and updated to reflect any changes. This proactive approach ensures the responsible handling of sensitive data and the continued effectiveness of public health interventions.

17 https://learn.cste.org/log-in
18 https://www.networkforphl.org/wp-content/uploads/2022/12/Network-for-Public-Health-Law-Data-Disaggregation-Handbook_FINAL.pdf, p. 43.

# Stage 4: Interpretation

This stage in the data release lifecycle centers on interpretations of the data being released for use. Its target outcome is to ensure that if interpretations are provided, they are inclusive and have minimal biases in facilitating data use, whether it be for decision-making, programming, planning, etc.

It is important to clarify that this stage may not involve making interpretations on behalf of requestors (e.g., stating what the data mean) but focuses instead on providing information on how to interpret the data responsibly. This includes highlighting relevant limitations and contextual factors that may affect interpretation. Providing this kind of clarity helps ensure transparency, accuracy, and equity in downstream data use.

To ensure the data interpretations are equitable and inclusive, it is essential to involve diverse perspectives in the review process. This section outlines the roles of internal epidemiologists and a data equity team, for each group's review process.

## Forms of Data Release

Aggregate data report interpretations are often developed by the epidemiologist and can take various forms depending on the audience and purpose.

**Interpretation may include:**

- **Narrative summaries:** Interpretation (developed by the epidemiologist) can include descriptive explanations of key findings.
- **Visual data representations:** Include charts, maps, tables within a dashboard. If the final product is intended for the public or media release, it can include a brief interpretation or key takeaways (developed by the epidemiologist) from the data to support understanding and informed use.
- **Aggregate data:** De-identified, aggregate-level data that are released for the purpose of analysis, reporting, or public dissemination often do not include data interpretation but should include the necessary information required for users to correctly interpret the data.
- **Web postings:** Data interpretations made accessible on public platforms should include both technical notes and plain-language explanations.

## Data Reliability and Quality

Clear communication on the reliability and quality of the aggregate data are essential to ensure informed interpretation and responsible use.

**Ensure the following are clearly communicated:**

- **Limitations** of the data (e.g., changes in case definitions, underreporting, or incomplete variables) should be explained.
- **Data sources** should be described, including the type (e.g., surveillance systems, vital records, electronic health records (EHRs), surveys), time frame of collection, and responsible agency.
- **Masking of small counts** <6 to protect privacy should be described and justified.
- **Contextual notes** on any analytic decisions made to address reliability issues (e.g., combining years or groups due to small sample sizes) should be clearly stated in plain language. When analysis or interpretation is not provided, these notes offer the context for requestors to make informed decisions about how to analyze or interpret the data.

## Review Process to Ensure Inclusive and Accurate Interpretation

### *Internal Epidemiologists' Review*

**Objective:** Ensure technical accuracy and methodological rigor in data interpretation while considering potential equity implications.

A review committee should be composed of lead/senior epidemiologists and data analysts from each department within the division. If a submission originates from their program area, they may be excluded from participating in the committee review process.

**Review Focus:**

- **Accuracy:** Verify the correctness of data analysis and interpretation. Ensure that statistical methods and results are sound.
- **Contextual understanding:** Assess how well the interpretation accounts for different population characteristics and potential disparities.
- **Bias identification:** Look for any inherent biases in the data analysis or interpretation process.
- **Terminology:** Ensure that terms such as correlation and causation, incidence and prevalence, and risk ratio and odds ratio are defined and explained at a reading level appropriate for the intended audience.

**Process:**

- **Initial assessment:** Conduct an internal review of data interpretation for methodological accuracy and relevance.
- **Documentation review:** Examine the documentation for clarity and completeness in describing the data sources, methods, and limitations.
- **Feedback integration:** Provide recommendations for addressing any identified issues and ensure that any necessary revisions are made.
- **Next steps:** After the Epi Review Team communicates its approval, additional review may be needed prior to final sign-off. Note: Media requests will likely have a different approval process.

4

### *Data Equity Team Review*

**Note:** If the infrastructure exists, make time to consult with a data equity focused health department team.

**Objective:** Evaluate the data visualizations, reports, outputs, and interpretation through the lens of equity, disparity, and inclusion, ensuring that data presentations are respectful, accurate, and responsive to the diverse needs, and identities of the communities represented.

**Review Focus:**

- Equity and disparity considerations:
  - Assess whether the data and their interpretation acknowledge and address disparities in outcomes, access, or opportunities between population groups.
  - Consider whether the analysis identifies root causes or contributing factors to observed disparities.
  - Examine how the findings may impact different communities differently—intentionally or unintentionally.
- Data inclusion and suppression practices:
  - Ensure that masking of small counts protects confidentiality without erasing representation.
  - If subgroups must be collapsed or suppressed, clearly identify them in the accompanying narrative to maintain visibility and accountability.
  - Evaluate whether data suppression may disproportionately affect people who have historically been underserved by state or local governments.
- Cultural sensitivity and relevance:
  - Assess whether the language, framing, and visuals used in reporting are culturally appropriate, respectful, and accessible.
  - Ensure community norms, values, and lived experiences are acknowledged and reflected in the data narrative.
- Bias awareness and inclusion framing:
  - Identify potential sources of bias in data collection, analysis, or presentation.
  - Recommend strategies to incorporate diverse perspectives, particularly those from underrepresented or directly affected communities.
  - Promote inclusive language and framing that avoids deficit-based or stigmatizing narratives.

4

**Process:**

- **Review:** Analyze the interpretation to ensure that it is aligned with data equity principles. This includes verifying that harmful narratives are not being perpetuated in the data interpretation, that person-first language is used, and that the interpretation adheres to standard language when discussing marginalized communities.
  - Example for person-first language: people experiencing homelessness, people with a disability, people living with diabetes, people who use drugs / people who inject drugs..
- **Feedback mechanism:** Engage with other teams that work directly with community members to collect insights and suggestions for improving the interpretation from an equity perspective.
- **Recommendation integration:** Adjust the interpretation based on feedback to better reflect equity considerations and address any gaps identified by the team.

### *Community Engagement Review*

**Note:** If the infrastructure exists and time permitting, allocate time to consult with the communities represented or impacted by the data.

**Objective:** Incorporate the perspectives of community representatives to ensure that the data output and interpretation are relevant, understandable, and respectful of the lived experiences of those affected.

**Review Focus:**

- **Community relevance:** Assess whether the data released accurately reflect the experiences and concerns of the community.
- **Transparency and understanding:** Ensure that the interpretation is communicated clearly, with definitions of terms, and that the interpretation notes are as accessible as the data are to the community.
- **Compensation and respect:** Ensure community representatives are appropriately compensated for their time and input.

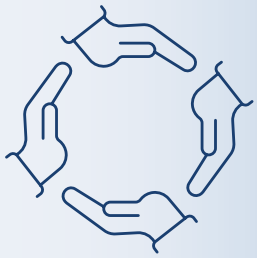
**Process:**

- **Focus on an effective engagement strategy.**
- **Consultation:** Reach out to community representatives to review reports, data visualizations, and interpretation, using methods such as focus groups, surveys, or interviews to gather their input.
- **Feedback collection:** Gather in-depth feedback on how the data interpretation reflects community experiences and priorities, ensuring that the information presented in reports and dashboards is relevant and meaningful to the community.
- **Revision based on feedback:** Revise the interpretation based on community feedback to ensure it is inclusive and respectful of their perspectives and reading levels.
- **Documentation:** Document how community feedback is used to adjust and improve interpretation.
- **Compensation:** Ensure fair compensation for community representatives to acknowledge their contributions and encourage meaningful participation.

4

## Data Equity Checkpoints

- Who was not represented? Refer to tools such as Northwest Center for Public Health Practice's "Building Inclusive Data Visualizations."[19]
  - Note: Interpretation could be from individuals without lived experience.
- Ensure equity and transparency in data interpretation.
- When applicable, be transparent about the interpretation process.
  - Review statistical methods and check for positional bias (i.e., bias that can result from the social, cultural, or professional perspective of the data source, methodology, or analyst that supports preexisting beliefs).
  - Be explicit about positionality (i.e., the acknowledgment of how an individual's background, values, identity, or institutional role may influence how data are interpreted or presented).
- Ensure the use of appropriate and person-first language.
- Name subgroups explicitly in the accompanying narrative or note when data are collapsed, to ensure the populations remain visible and acknowledged.
- Evaluate interpretations for cultural relevance and appropriateness to avoid reinforcing harmful stereotypes or narratives.
- Assess whether equity issues are addressed or overlooked and consider differing impacts across communities.
- Identify and mitigate potential biases, recommending strategies to include diverse perspectives.
- Engage with community representatives and incorporate their perspectives meaningfully, ensuring they are fairly compensated, acknowledge their contributions and support sustained, equitable participation.



## Legal Considerations Checkpoints

- It is important to understand, from your legal counsel, how interpretation is allowed to be provided (written versus verbal) by the health department. When interpretation is appropriate and allowed, it can help provide meaning to the data.

19 https://www.nwcphp.org/training/inclusive-data-visualizations

# Stage 5: Approval

This stage in the data release lifecycle centers on approval. Formal approval steps are important to ensure that data requests are managed consistently, efficiently, and transparently. Key steps include verifying the request's scope and suitability, assigning it to the right program, and addressing gaps in data or jurisdictional practices through iterative review processes. Programmatic approvals assess whether data meet established practices, including analysis and interpretation. The final approval involves ensuring all steps are complete, assigning roles for dissemination, and engaging communication partners to facilitate effective release. There are multiple points during the data release lifecycle that may benefit from a progress check and approval.

Below is a structured chronology of the key approval steps and decision points during data request processing. Some of these steps will likely occur earlier in the overall data release lifecycle. Having a clear protocol for staff to get approval is essential.

## 1 Approval of Initial Receipt of Request

- **Consider requestor type:** Does the type of requestor change the approval process?
  - Determine if different or additional approval processes or criteria are needed based on whether the requestor is internal, external, a member of the public, a government agency, etc.
  - Having separate workflow and approval checklists based on requestor type may be useful.
- **Determine if the request meets the scope and assign the request:**
  - **If no:** Respond to the requestor, explaining that the submission doesn't meet the scope (e.g., it is not a data request, it lacks essential information, etc.). Suggest alternative ways to request the data or alternative data sources, if applicable.
  - **If yes:** Assign the request to a specific program or person responsible for guiding it through the agency's processes (approve for further review and consideration).
- **Consider appropriateness of specified program:** Does the specified program accept assignments for request fulfillment?
  - **If no:** Reassign the request to an appropriate program or person.
  - **If yes:** Initiate the agency's procedures for reviewing and fulfilling the data request and mark the assignment as approved.

## 2 Programmatic Approval of Requestor Expectations and Fulfillment of Request

- Clarify expectations:
  - Is further discussion needed?
    - **If no:** Approve the current expectations and proceed.
    - **If yes:** Reach a consensus with the requestor on the expected data product before approving further steps to compile the data.
- Assess the compiled data:
  - Do the data meet expectations?
    - **If no:** Document any gaps and reasons for incomplete data. Consider modifying expectations (e.g., subdividing the request for partial or sequential fulfillment).
    - **If yes:** Approve the raw data.
- Evaluate analyses and interpretation:
  - Do analyses and interpretations meet expectations?
    - **If no:** Document gaps and reasons for incomplete supplemental information related to the data.
    - **If yes:** Approve the analyses and interpretations.

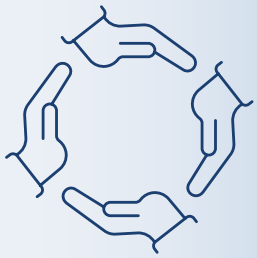
## 3 Final Approval for Dissemination or Release of Final Data Product(s)

- **Determine who gives final approval:** Identify the specific roles or individuals responsible for giving the final approval. This could be:
  - A senior epidemiologist, data officer, or equivalent within the agency
  - The director of the program or bureau handling the request
  - A data governance committee or board
- **Review the fulfillment steps and programmatic approvals:**
  - Ensure that the fulfillment process is completed. A checklist could be a helpful reference to make sure that all steps are complete.

This structured approach ensures that all necessary checks and approvals are made at each step of the data request process, facilitating clear communication, transparency, and consistency.

## Data Equity Checkpoints

- Consider convening an equity review board that is consulted for overall incorporation of equity.
- During approval, ask, "Does analysis or data product meet the expectations for equity considerations?"
- Consider vetting the information or requesting approval from the groups represented in the data release.



5

## Legal Considerations Checkpoints

- To ensure a consistent, transparent, and legally sound internal review process, agencies should implement clear guidance on when and how to engage legal or compliance experts during the data release lifecycle.
- Legal review should not be an afterthought or final sign-off only. Instead, consult legal counsel early and proactively. Keep them informed during iterative programmatic reviews.
- Ensure all reviewers (e.g., legal, programmatic, communications) are working from the same understanding of the data's intended use, sensitivity, and scope.
- Build ongoing relationships between program teams and legal counsel so that the legal guidance is practical and not just procedural.
- Encourage a culture where program staff see legal review as supportive rather than restrictive


# Stage 6: Release/Dissemination

The last stage in the data release lifecycle centers on having a dissemination or communication plan for the data that are being released. Its target outcome is to ensure the final report or dataset is effectively communicated and reaches its intended audience with the necessary context for appropriate use.

When creating a dissemination or communication plan to fulfill an individual aggregate data request, consider the following key components: information to accompany the aggregate data release, interpretation of the aggregate data, audience awareness, publicity and communication methods, and aggregate data release tracking, all of which are detailed below.

## Information to Accompany Data Report

- **Accessibility:** Ensure the aggregate data are accessible to all intended audiences, including those using mobile devices. Anticipate new accessibility guidelines (such as WCAG 2.2[20]) and align with these standards.
- **Language:** Incorporate plain language[21] principles to facilitate the dissemination and communication of clear public health information. Provide translations in priority languages, keeping in mind potential delays or additional funding requirements for translation services.
- **Interpretation guidelines:** Include guidance on how the data should be used or interpreted, especially if the findings have public health implications.
- **Build data literacy in the community:** Offer resources or training to help users understand and appropriately interpret the data. Collaborate with the requestor and external allied partners such as colleges, libraries, and community organizations to grow health data literacy. Actionable Intelligence for Social Policy (AISP) suggests using a range—from light engagement through public activities like data "gallery walks" to more intense involvement, such as community-based participatory action research.

20 https://www.w3.org/WAI/standards-guidelines/wcag/
21 https://www.cdc.gov/health-literacy/php/develop-materials/plain-language.html


6

## Interpretation of Data

- **Provide clear descriptions of the data,** such as table titles, data source, date produced, definitions, methodologies, limitations, and any standardized footnotes regarding provisional data, data owners, funding source, and accuracy or completeness, to assist with interpretation.
- For public-facing dashboards or reports, **consider including a simplified interpretation** of what the data show through executive summaries, concise bullet points, and up to five key takeaways. Including plain language summaries helps provide context to the public and other partners who may not have a technical background so they can better understand the data being presented.
- **Encourage presentation of data using a storytelling framework** to explain and highlight why the data are important in answering a question or prompting a response or action. Various universities and public health organizations, such as WHO, CDC, and CSTE, offer tools and training to improve the impact and efficacy of public health communication through storytelling (see References & Resources for storytelling aids).

## Audience Awareness

- **Identify who needs to be aware** of the aggregate data release, including both internal partners (e.g., public information officers, program managers) and external partners (e.g., specific public groups with interest in the data).
- **Plan for targeted outreach** to these groups, using different communication channels as appropriate. Keep in mind that partners may have additional resources and communication channels available to disseminate the data.

## Publicity and Communication Methods

- **Press releases:** Use press releases for major data releases, such as new reports or dashboards. Ensure the relevant teams, such as communications departments, are prepared and equipped to disseminate these releases effectively.
- **Type of release:** Decide when to use static reports versus interactive dashboards based on the data complexity and the audience's needs.
- **Social media or other amplification methods:** Consider the use of social media to share data and garner attention to the findings.[22]

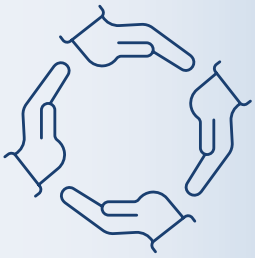
## Data Report Tracking

Implementation of a tracking process for all data releases, both to specific parties and the general public, was introduced during the intake stage. Use of a tracking system helps ensure transparency, accountability, and follow-up if necessary. The tracking record may be maintained and updated throughout the data release lifecycle. Release and dissemination of aggregate data products are typically associated with review of a request-specific tracking record for completeness and accuracy, as well as final record close-out.

22 Kanchan S, Gaidhane A. *Social Media Role and Its Impact on Public Health: A Narrative Review*. Cureus. 2023 Jan 13;15(1):e33737. doi: 10.7759/cureus.33737. PMID: 36793805; PMCID: PMC9925030.

## Data Equity Checkpoints

- **Accessibility needs:** Electronic data visualizations and documents should be ADA accessible.[23]
- **Communication source:** Consider whether the requested information would be more effectively received by the audience if communicated by another group or agency—particularly when co-developing materials for external dissemination or when specific groups may be impacted.
- **Should the data be disseminated and to whom:** Tailor data release to intended audiences with consideration for unintended audiences and steps to protect against unintentional/unintended uses.
- **Requestor perspective:** Allow requestors to interpret from their experience and perspectives. Encourage collaboration with impacted populations or communities to effectively communicate to audiences.
- **Resources on equity are found in the References & Resources section.**



## Example Checklist for Preparing Data for Release

- ☑ **Jurisdiction check:** Verify the appropriate dataset, subset, and jurisdiction (e.g., state residents versus specific occurrences).
- ☑ **Data compliance:** Ensure compliance with the agency's data release policy and appropriate presentation of tables.
- ☑ **Data description:** Use descriptive titles, footnotes, and indications of data versions, provisional data, and any limitations.

This comprehensive approach ensures that data are effectively communicated and reach their intended audience with the necessary context for appropriate use.





23 Guidance and tools for creating ADA accessible content are available at https://www.ada.gov/resources/effective-communication/.

# Appendix A: Summary of Learnings from External Consultant-Led Data Privacy Hub Interviews and Focus Groups

## Summary of Key Meetings and Topics Discussed

**Three key meetings were held with the members of DRPASS on:**

- September 26, 2023
- October 10, 2023
- October 31, 2023

During these meetings, several critical topics were discussed in small breakout rooms. These topics included existing policies/procedures, challenges with working with data (small cell sizes, data sensitivity), community engagement and data reporting, analysis, and data protection techniques (suppression, statistical disclosure control, aggregation strategies, etc.). In addition to workgroup breakout discussions, the privacy consultants held additional focus groups to inform future privacy training development for CSTE.

In winter 2023, focus groups were held. The following points were consistently raised during these discussions and, therefore, deemed by the privacy consultant to be key areas that the training material should cover:

- **Principles of re-identification risk:** Multiple focus group participants noted that, as a starting point, a basic overview of HIPAA regulations should be provided and that it would be beneficial to state what is meant by the risk of re-identifying individuals. Some participants noted that, due to the open-ended wording used within HIPAA, there can be multiple interpretations of the guidance (both internally to teams/departments and externally between local/state guidelines). An overview would help provide a solid foundation for any subsequent remediations discussed as part of the training.
- **Geography in relation to re-identification risk:** Geographical constraints on data reporting were frequently mentioned during the focus groups. For geographically isolated populations, there is often a limit on the granularity of information that can be revealed, particularly in cases where only a single facility exists in the region. This can have an impact on the accurate representation of communities in the final, compliant dataset. In particular, it was noted that many Tribal health regions do not align with ZIP codes and, therefore, resolving geographical areas to align with statewide reporting requirements is often challenging and frequently does not accurately represent the Tribal communities.
- **Balancing equity and privacy:** Constraints on the granularity of reporting for Tribal communities, as well as other small population groups (e.g., geographically isolated), often result in inaccurate representation of these communities or, in some cases, the population groups not being reported at all in public data. It was noted that it is common for these communities to be aggregated with other areas or to have "noise" (i.e., erroneous data points to distort analysis) added to the data, which ultimately can result in the health status of a community being incorrectly captured. One example given was for blood pressure monitoring, where populations that were at risk of high blood pressure were not being properly represented due to the risk of re-identification (e.g., Native American communities). Similarly, it was noted that vaccination rates against COVID-19 are unknown to this day for some

Tribal communities due to the lack of transparent reporting. Focus group members requested guidance on how best to deal with situations like this, where privacy and equity seem to be at odds.

- **Developing internal guidance:** While some focus group participants noted that their departments had comprehensive data release policies, others noted that there was a lack of consistent internal guidance to which they could refer. Advice on best practices and any potential methods for reporting standardization was requested. Additionally, guidance on any methodologies for automating the processes of standardization or suppression was identified as a need.

## Privacy Evaluation/Suppression Needed

The group discussed how the community can sometimes inadvertently risk individual privacy. One example provided detail on how a hospital had inadvertently released data, which the public health epidemiologists then had to substantiate. While nobody in the group reported that their surveillance reporting had caused any data breaches, they acknowledged that these types of incidents present a privacy risk. The mitigation approach in the above example was to provide data for a year for that county.

Another problem to consider is that social media can inadvertently identify previously unknown individuals in epidemiological data. For example, a COVID-19 case reported at a school could be identified by people on social media who attend or live near the school in question. One participant noted that there is a very small town in her state that has only 65 people, where it is highly likely that individuals know a significant amount of information about other individuals in this community.

One aspect of community engagement that presents challenges for epidemiologists is how to handle media inquiries. Guidance on reporting data at various levels of granularity would be helpful. Media entities sometimes seek corroboration about cases and want transparency from health departments. Balancing the media's requests for transparency with individual privacy is difficult, especially given the media's role in getting information into the hands of the community so people can take personal protective measures (e.g., COVID-19).

In a similar vein, participants said the COVID-19 pandemic increased the demand that key data takeaways, as well as the underlying data that drove those findings, be highlighted clearly. These high-level findings are more easily digestible for the public and elected officials and can quickly aid decision-making. Visual representations of data are often more effective, as they provide a clearer and more easily interpretable view.

**Some key findings from the discussions were as follows:**

- Participants varied on the minimum number of individuals in a cell that is required to exceed suppression thresholds. Some suppress cells with fewer than five people or provide a confidence interval for rates. One site doesn't report deaths in cells of fewer than three individuals, and another doesn't report on denominators lower than 50,000.
- Some have difficulty with reporting on small populations in a huge geography and have developed custom geographies for reporting purposes.
- In relation to de-identification of data in accordance with the HIPAA Privacy Rule, some participants stated that they use the Safe Harbor method for de-identifying data, while others rely on the expert determination method. Expert determination was noted as being done in-house and by leveraging an external expert. One member acknowledged that use of an external expert declined after the member worked to put strong internal guidelines in place.

Based on the points raised by the discussion group, the following training areas were identified:

- Guidance on data suppression
  - Methodologies for primary and secondary suppression and any potential automation that can be applied to these processes
  - Numerator and denominator-based suppression methodologies
  - Reuse of data suppression algorithms; understanding whether algorithms can be used continuously or if they require modification between releases/on a defined regular cadence
  - Preventing calculations through backward analyses
  - Implementation of suppression for real-time data reporting (e.g., live dashboards)
  - Methodologies for data masking as a suppression tool
  - Understanding the effects of a changing population on suppression methodologies
- Understanding HIPAA Privacy Rule
  - Comparison between the Safe Harbor and expert determination approaches
  - Guidance on factors affecting risk in datasets
  - Thresholds for re-identification risk

## Data Suppression Approaches and Privacy

Based on the points raised by the discussion group, the following training areas were identified:

- Explanation of different methods of data de-identification and how these can be utilized to balance equity and privacy
- Reporting of specific populations of interest (e.g., based on race/ethnicity, disease area)
- Reporting with population size considerations
- Situational awareness; understanding the impact of data reporting on communities

## Data Reliability and Quality

**Note from consultant on guidelines or SOPs:** The group agreed that any guidelines or standard operating procedures (SOPs) need to have the following characteristics:

- They need to be consistently interpretable.
- They need to allow for replicable results.
- They must account for varying sensitivity of data being reported.
- They need to account for the varying levels of statistical sophistication required to respond to requests.
- They need to consider the conditions under which the data are being reported (community, political, etc.).
- They need to be understood by the audience. Departments struggle with explaining data privacy needs to the public. (e.g., could be helpful to include flow charts)

# Appendix B: Common Statistical Disclosure Control Techniques & Steps for Implementing SDC

Statistical disclosure control (SDC) is essential for responsibly managing and sharing data in public health. By applying a range of SDC techniques, jurisdictions can protect individual privacy, comply with legal requirements, and ensure that data remain useful for decision-making. The challenge lies in finding the right balance between minimizing disclosure risk and preserving data utility, which requires careful planning, ongoing evaluation, and partner collaboration.

## Common Statistical Disclosure Control Techniques

1. Data Suppression
   - Suppress or remove data elements that pose a high risk of disclosure, such as small counts or sensitive attributes. For example, replacing small cell counts (e.g., fewer than five individuals) with symbols like "<5."
   - Use suppression for both direct identifiers (such as names or Social Security numbers) and indirect identifiers (such as unique combinations of demographic variables).
2. Data Aggregation
   - Combine categories or groups to reduce the level of detail in the data for numerators or denominators. For example, instead of providing age in exact years, use age ranges (e.g., "0-4," "5-9") or combine smaller geographic areas into larger regions (e.g., combining multiple ZIP codes or counties into districts) to reduce the risk of identifying individuals in small populations.
3. Top-Coding and Bottom-Coding
   - Limit the granularity of extreme values in a dataset. For example, age above a certain threshold might be grouped into a single category (e.g., "over 65") to avoid revealing exact age.
4. Data Masking and Redaction
   - Mask or redact sensitive data elements, such as replacing names or addresses with generic labels (e.g., "Person A" or "Location X").
5. k-Anonymity and How to Address Limitations
   - **k-anonymity:** A powerful method for protecting individual privacy by ensuring that no individual can be uniquely identified in a dataset based on personal attributes. K-anonymity requires that each equivalence class contains at least k records. So, if k=2, this means that for every set of quasi-identifying attributes there must be at least two records with the same attribute values. This prevents any individual from being identified uniquely by a combination of quasi-identifiers.
     - **Example of k-anonymity:** Suppose you have a dataset with attributes like age, gender, and ZIP code. If there are unique combinations of these attributes (e.g., a 30-year-old female in ZIP code 12345), k-anonymity might not be satisfied. To achieve k=3, the data could be modified so that age is represented as a range (e.g., "30-35 years") and ZIP codes are generalized to the first three digits (e.g., "123xx"). Now, at least three individuals share the same generalized values for age, gender, and ZIP code, making it harder to identify any one person.

- **Limitations of k-anonymity:** To address the limitations of k-anonymity, more advanced techniques, such as **l-diversity** (which ensures diversity of sensitive attributes within each equivalence class) and **t-closeness** (which ensures the distribution of sensitive attributes within an equivalence class is like the overall distribution), have been developed.

## Steps for Implementing Statistical Disclosure Control (SDC)

1. Assess Disclosure Risk
   - Evaluate the potential risk of disclosing sensitive information from the dataset. Consider factors like the uniqueness of records, the size of the dataset, the availability of external data sources, and the sensitivity of the data.
2. Choose Appropriate SDC Methods
   - Implement the chosen SDC techniques on the dataset, carefully documenting the steps taken to modify the data and the rationale behind each technique.
3. Apply SDC Techniques
   - Implement the chosen SDC techniques on the dataset, carefully documenting the steps taken to modify the data and the rationale behind each technique.
4. Evaluate Data Utility
   - After applying SDC methods, assess the impact on data utility. Ensure that the modified data remain useful for their intended analytical purposes while protecting privacy.
5. Iterate and Optimize
   - If necessary, adjust the SDC techniques to achieve a better balance between privacy protection and data utility. This may involve reapplying or modifying methods to reduce disclosure risk without unduly compromising the dataset's usefulness.
6. Document the Process
   - Maintain thorough documentation of the SDC process, including the methods used, their impact on the data, and any limitations. This ensures transparency and allows others to understand how the data were protected.

## Best Practices for Implementing Statistical Disclosure Control

- **Transparency:** Clearly communicate any changes made to the data, including the methods used and their potential impact on data interpretation.
- **Regular review:** Continuously review and update SDC methods to address new risks or changes in the external data environment.
- **Partner engagement:** Involve partners (e.g., data users, privacy experts, community representatives) in discussions about SDC methods and their trade-offs to build trust and ensure the data meet their needs.

# References & Resources

## Health equity resources


**Actionable Intelligence for Social Policy (AISP) A Toolkit for Centering Racial Equity Throughout Data Integration.** https://aisp.upenn.edu/wp-content/uploads/2022/07/AISP-Toolkit_5.27.20.pdf

**Islands Health Equity Framework.** ASTHO. https://www.astho.org/topic/resource/island-equity-framework/

**Preferred Terms for Select Population Groups & Communities.** CDC Restored 2025. https://restoredcdc.org/www.cdc.gov/health-communication/php/toolkit/preferred-terms.html

**We All Count.** https://weallcount.com/the-data-process/


## Storytelling resources


**Kanchan S, Gaidhane A.** Social Media Role and Its Impact on Public Health: A Narrative Review. Cureus. 2023 Jan 13;15(1):e33737. doi: 10.7759/cureus.33737. PMID: 36793805; PMCID: PMC9925030. https://online.hbs.edu/blog/post/data-storytelling

**Region IV Public Health Training Center: Data Visualization and Storytelling** https://r4phtc.org/data-visualization-and-storytelling-for-public-health-professionals-how-to-present-your-data-in-meaningful-and-impactful-ways/

**WHO Communicate for Health:** Storytelling https://www.who.int/westernpacific/initiatives/Communication-for-Health-C4H/Storytelling#


## Report release/dissemination equity resources


**A Toolkit for Centering Racial Equity Throughout Data Integration.** Developed by Actionable Intelligence for Social Policy. https://assets.aecf.org/m/resourcedoc/aisp-atoolkitforcenteringracialequity-2020.pdf

**Do No Harm Guide: Applying Equity Awareness in Data Visualization.** Developed by Urban Institute. https://www.urban.org/sites/default/files/publication/104296/do-no-harm-guide.pdf

**Public Health Data Disclosure or Request Readiness Assessment. ASTHO 2025.** https://www.astho.org/topic/resource/public-health-data-disclosure-or-request-readiness-assessment/

**Transforming Public Health Data Systems.** Robert Woods Johnson Foundation. https://www.rwjf.org/en/insights/our-research/2021/09/transforming-public-health-data-systems.html

## Example dashboard resources

**Health equity dashboard.** Orange County, New York. https://orangecountynydoh.shinyapps.io/healthequitydashboard/

**Bright Lights, Bold Ideas.** NACCHO 360. https://naccho2024.mapyourshow.com/8_0/sessions/session-details.cfm?scheduleid=118

**Kentucky Health Equity Dashboard.** https://dashboard.chfs.ky.gov/views/KentuckyHealthEquityDashboard/KPMMain?%3Aembed=y&%3AisGuestRedirectFromVizportal=y

**Developing a Data Dashboard to Address Health Equity Concerns: Insights from Puerto Rico.** ASTHO. https://www.astho.org/topic/resource/developing-data-dashboard-address-health-equity-concerns-insights-puerto-rico/