Guidance Document on Creating and Releasing Hospital and Facility Discharge Data Public Use Files

January 2012

The National Association of
Health Data Organizations

Guidance Document on Creating and Releasing Hospital and Facility Discharge Data Public Use Files

Acknowledgements

NAHDO extends a special thanks to the NAHDO State Data Release Guidelines Workgroup who contributed to the development of this document.

- **Sallie Milam**, JD, CIPP, CIPP/G, West Virginia's Chief Privacy Officer, West Virginia Health Care Authority
- **Joseph Greenway**, MPH, Director, Center for Health Information Analysis, University of Las Vegas, Nevada
- **Michael A. Stoto**, PhD, Professor of Health Systems Administration and Population Health, Georgetown University
- **John G. Morgan**, Hospital Discharge Database Manager, Utah Deptartment of Health
- **Carol Stocks**, Program Analyst, Agency For Healthcare Research and Quality
- **Amy B. Wenmoth**, MA, Director of Clinical Analysis, West Virginia Health Care Authority
- **Jessica L. Toth**, MSPH, Epidemiologist, West Virginia Health Care Authority, Clinical Analysis
- **Susan Schow**, MPH, Epidemiologist, Maine Health Data Organization
- **Doris J. Green**, Health Program Specialist, Hospital Discharge Data Section, Arkansas Department of Health

Document Authorship

- **Denise Love**, RN, MBA, Executive Director, NAHDO
- **Emily Sullivan**, MS, MPH, Director of Projects, NAHDO
- **Barbara Rudolph**, PhD, MSSW, Consultant to NAHDO, and Senior Scientist, University of Wisconsin—Madison .

Table of Contents

Guidance Document on Creating and Releasing Hospital and Facility Discharge Data Public Use Files

Executive Summary

Almost every state has some form of a hospital discharge data reporting system, maintained by a state or private Health Data Organization (HDO). These data systems capture all hospitalizations for every patient, regardless of insurance status or payer. As the demand for healthcare data increases, HDOs are seeking guidance on suitable standards for data release practices. Given the significant role of these healthcare data systems for healthcare reform, state budgeting, and the need for patient consumer information, health data organizations must be as transparent as possible in how they collect and release data, without doing harm or diminishing patient privacy.

HDOs must balance competing perspectives on data release. Because of their unique missions, state health data organizations have been innovators in efficient and effective data dissemination solutions, providing large volumes of hospital discharge data for diverse users and uses. HDOs typically release a range of data reports and files which can be depicted along a continuum from lower risk to higher risk of re-identification of a patient, depending on the type of file and elements. Because of their utility for many applications, Public Use Files (PUFs) are produced and released by most HDOs. They contain micro-level data elements describing inpatient stays and discharges from hospitals, and retain important hospital financial and utilization information from the patient's stay, but remove all patient identifying information, such as name, home address, social security number, and birth date.

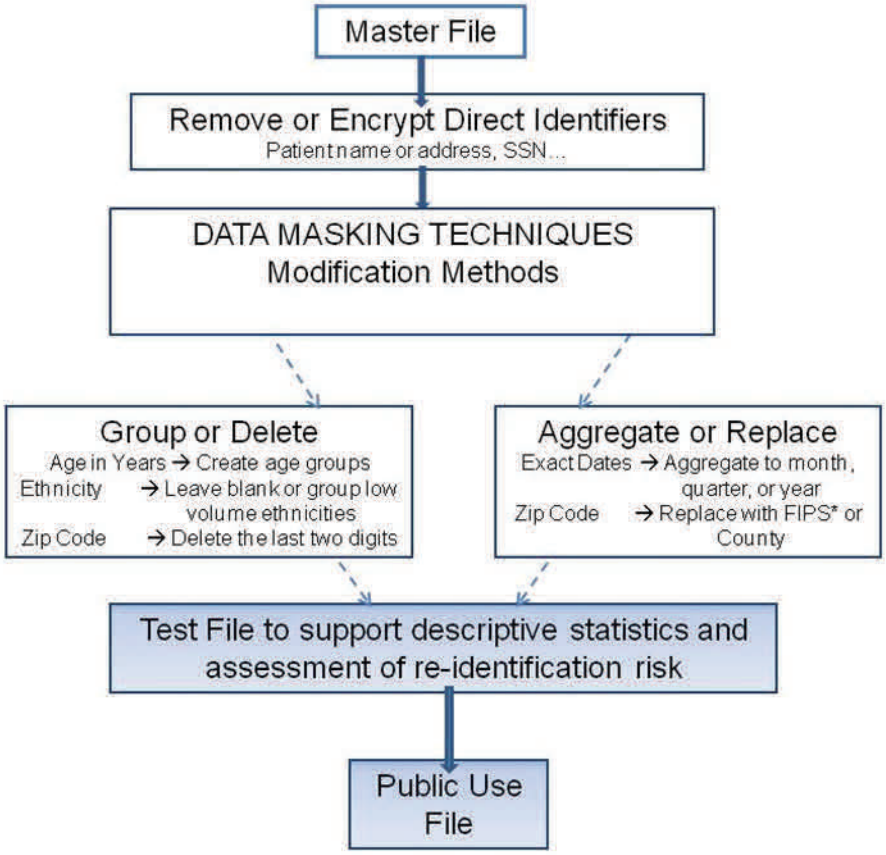
The National Association of Health Data Organizations (NAHDO) in collaboration with the State Data Release Guidelines Workgroup produced this document to guide HDOs and others who are seeking guidance on best practices for disseminating hospital discharge data files and the privacy parameters for such releases. Specifically this guidance document aims to promote the release of useful information in PUF creation for multiple users and uses by:

- Informing and guiding data release policies and practices in state health data organizations;
- Providing evidence-based practices to anonymize hospital data sets in order to protect the privacy and confidentiality of the patient.

As part of this project, NAHDO reviewed Public Use Files from 18 state HDOs. Common elements released in state PUFs can be categorized into the following groups;

- *Non Restricted Data Elements (LOW CONTROL).* Data elements that can be released without further modification as they are not considered alone or in combination as identifiable of the individual patient. Typically include utilization data elements and facility data elements.
- *Restricted Data Elements (MODERATE CONTROL).* Elements in this group are typically released in a PUF but are often modified when a combination of these data elements creates a unique record. Generally these include patient demographic elements.
- *Enhanced PUF elements.* Data elements may not directly or indirectly identify patients but are not commonly included in state PUFs. These data elements such as a transfer or readmission flag can increase the utility of the file.

The figure below outlines the common statistical approaches/methodologies the workgroup identified and agreed on as useful for a state creating a PUF.


Master File
Remove or Encrypt Direct Identifiers
Patient name or address, SSN…
DATA MASKING TECHNIQUES
Modification Methods
Group or Delete
Age in Years → Create age groups
Ethnicity → Leave blank or group low volume ethnicities
Zip Code → Delete the last two digits
Aggregate or Replace
Exact Dates → Aggregate to month, quarter, or year
Zip Code → Replace with FIPS* or County
Test File to support descriptive statistics and assessment of re-identification risk
Public Use File

In summary, the Workgroup acknowledges that the release of data in a PUF represents a series of trade-offs. These trade-offs include statistical and management controls deployed to reduce the risk of re-identification, while preserving as much data utility as feasible. More statistical restrictions diminish the utility of the data set while more restrictive management/institutional controls may permit the retention of more granular and useful data. Thus, it is important to note that agencies that apply multiple controls are more likely to release more detailed data.

Introduction

Given the significant role of healthcare databases and information in health reform, state budgeting, and the need for patient consumer information, health data organizations (HDOs) must be as transparent as possible in how they collect and release data, without doing harm or diminishing patient privacy. State health data organizations and others are seeking guidance on best practices for disseminating hospital discharge data files as well as seeking guidance on privacy parameters for such releases.

Almost every state has some form of a hospital discharge data reporting system that captures all hospitalizations for every patient, regardless of insurance status or payer. These data systems are maintained by the state health data organization which is charged with dissemination of information for multiple uses and users, including healthcare purchasers, local public health officials, health services researchers, policy makers, consumers, and federal and state agencies. The discharge data systems were designed to provide important information on healthcare utilization, healthcare disparities, disease surveillance, market analysis, efficiency of care, etc. State data organizations must balance competing perspectives on data release. They must continually demonstrate the worth of the information they maintain to taxpayers and elected officials, yet they are chronically under-funded, especially for analytic services. Data release policies that are too restrictive can reduce or eliminate the utility of the data; however, when a data system fails to provide meaningful information, the purpose of data collection is challenged and could result in budget cuts that effectively eliminate the system.

State Public Use Files:

- Are produced and released by state health data organizations. These Public Use Files (PUFs) are a significant component of their provision of meaningful information; they contain micro-level data elements describing inpatient stays and discharges from hospitals.
- Retain important hospital financial and utilization information from the patient stay, but remove all patient identifying information, such as name, home address, social security number, and birth date.
- May not meet the needs of some researchers that require more information than what is contained in a PUF, but when a PUF suffices, it saves the significant work needed to prepare a special research file for each user.

About This Document

This document was created by the National Association of Health Data Organizations (NAHDO) in collaboration with its State Data Release Guidelines Workgroup (see acknowledgements). NAHDO was approached by several of its state members seeking information about state hospital discharge data release policies and practices. NAHDO formed a State Data Release Guidelines Workgroup to guide the development of this document that describes common practices for publicly releasing hospital discharge data.

States commonly (but are not required to) release data in the following ways: 1). Aggregated Tables and Published Reports; 2). Public Use Files; 3). Custom Data Files ; 4). Research Data Files and; 5). Intra Agency/Public Health Specific Release (State Health Data Organization Dissemination Practices below describes each of these types of release products in detail). The Workgroup determined the scope of this document was to provide guidance on Public Use Files (PUFs) only. The guidance aims to promote the release of useful information (in PUFs) for multiple users and uses by:

- Informing and guiding data release policies and practices in state health data organizations;
- Providing evidence-based practices to anonymize hospital data sets in order to protect the privacy and confidentiality of the patient.

State HDOs can use this document to review and update their current policies and practices. This guidance document thus provides: a synthesis of best practices on data release in PUFs and state practices that protect privacy via data element groupings, categorizations and other masking practices; and, some guidance in determining what additional steps states have in place to protect patient privacy when the population of interest is small. States have also requested additional information on data release practices for research use, data releases to public health entities for surveillance purposes, and use of discharge data on public websites. Other areas of interest include evaluation of potential threats to privacy via data linkage. The request for more information on research files and data linkage issues will not be addressed in this document, but could be addressed in additional guidance documents in the future.

Methodology

The NAHDO Data Release Guidelines Workgroup reviewed and synthesized current best practices by states in regard to PUFs. This group met weekly over a 12 week time period. During that time the work group:

- reviewed data elements collected by states and released in PUFs;
- reviewed state data release laws/policies to understand practices;
- reviewed state data groupings and aggregation processes and protocols;
- reviewed cell size protocols focused on release of data in a PUF and other management controls that protect individual patients from identification;
- reviewed best components of data use agreements;
- identified the decision steps states use to create a public use data set that takes into consideration state legal provisions and common practices used by states to aggregate, suppress, modify, and control the release of micro-data files; and;
- assisted NAHDO in synthesizing the discussions into a draft guidance document.

The Workgroup determined that terminology regarding the release of these PUF datasets varied within and across states. Rather than develop a glossary of terminology, the Workgroup suggested that NAHDO embed working definitions within the paper. Attempting to come to a single definition for these terms would be counterproductive given the complexity in state laws.

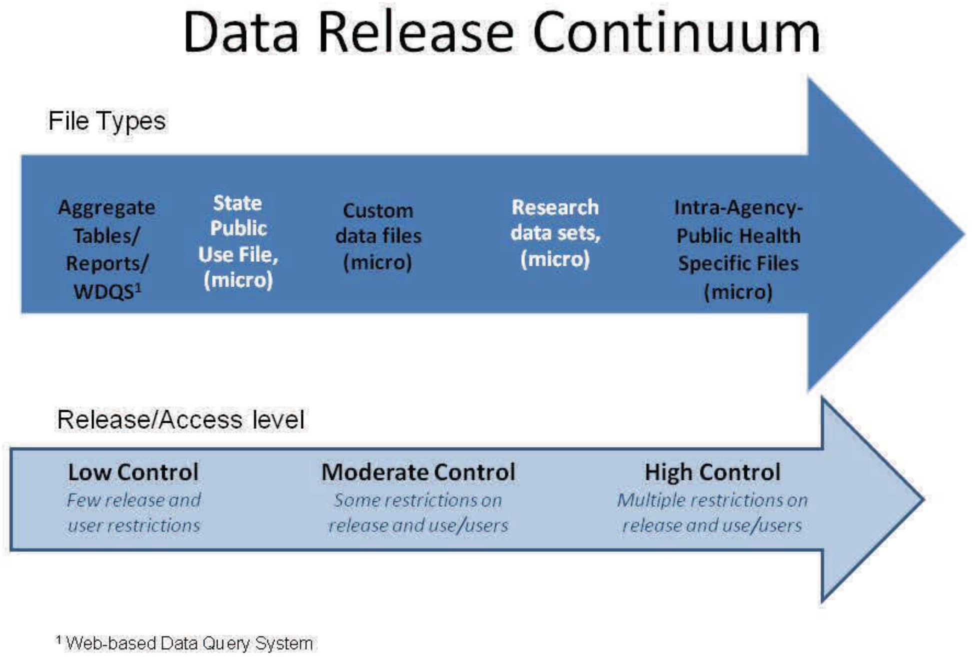
As part of the development process for this guidance document, a summary document was presented and reviewed at the national NAHDO conference in November of 2011. It is anticipated that a consensus on best practices by states will assist those states working on revisions and new rules and move states toward alignment with other leading state practices. In addition, a number of states provided comments on the final draft of this paper.

State Health Data Organization Dissemination Practices

State HDOs have been innovators in efficient and effective data dissemination solutions, providing large volumes of hospital discharge data for diverse users and uses. Hospital discharge data contain information about individual patients as well as financial and clinical aspects of hospital care; therefore the release of this information must carefully balance the risk of inappropriate disclosure with the public good that this information provides. To manage the risk of inappropriate uses of the data from external parties attaining potentially sensitive information about a patient's hospital stay, state health data agencies typically create a range of information products tailored to meet multiple user needs, while protecting patient information.

These data products can be depicted along a continuum from lower risk to higher risk of re-identification of a patient, depending on the type of file and elements within the file as well as the level of controlled access. Figure 1, Data Release Continuum below, illustrates that as data products retain more detailed information the corresponding controls and protections governing the data increases. The risk of patient re-identification increases as one shifts from aggregated tables and statistics to micro-data files. The most restricted of all data sets are interagency data repositories or warehouses, with limited authorization for access to a few qualified individuals. All of these data products are carefully designed by the state HDOs to meet the needs of various user groups and to carefully balance the trade-off between value of the information for the public good and protection of patient information. The PUF is categorized in the moderate risk range, as these files are de-identified through statistical modifications. Some states do permit the release of research-oriented data sets which may contain more specific patient information, and have additional controls and restrictions surrounding the data sets release and uses.

Figure 1: Data Release Continuum


Data Release Continuum
File Types
Aggregate Tables/ Reports/ WDQS[1]
State Public Use File, (micro)
Custom data files (micro)
Research data sets, (micro)
Intra-Agency-Public Health Specific Files (micro)
Release/Access level
Low Control
Few release and user restrictions
Moderate Control
Some restrictions on release and use/users
High Control
Multiple restrictions on release and use/users
[1] Web-based Data Query System

The following section describes the various data products/files typically produced and released by a state HDO as it relates to Figure 1. While data product content may vary across states, most release one or more of the following types of information.

Aggregated Tables and Published Reports

The least risk is associated with aggregated tables and published reports. These reports, and web-based data query systems, contain statistical information derived from the underlying databases. Agencies still must assess whether an individual could be identified within the aggregated data by an end-user; if there is a perceived threat of identification, states' apply masking techniques or remove from display those statistics where the aggregation shows there are only a few subjects. Many states now sponsor sophisticated Web-Based Data Query Systems (WDQS) for both public use and for secure public health networks. Because the database remains protected behind a secure firewall, the release risk is minimal; however, because queries are dynamic, the agency must employ sophisticated methods to assure the suppression of certain statistical configurations. Readers are referred to NAHDO's technical paper series, *"Public Health Data Dissemination Guidelines"* [1] and to published papers on State Web-Based Data Query Systems (WDQS) in the 2006 Journal of Public Health Management and Practice, volume 12 (2), March-April 2006.

Public Use Files

A data set that is in wide use across states and which has a moderate level of risk, if properly de-identified, is the Public Use File (PUF). PUFs have been in use successfully by state health data agencies for thirty years without inappropriately releasing information that could identify individuals. The de-identification process in the production of a PUF involves the removal and/or modification of patient identifiers that may otherwise allow an individual to be identified. This process of de-identification is described in a later section called "Common De-Identification/Statistical Modification Techniques of Patient Demographic Data Elements".

(Note: State definitions of de-identification may vary from the HIPAA definition of de-identification.)

The Workgroup determined that the scope of the guidelines should initially focus on PUF practices. The PUF is the most common data product released by state health data organizations, and the disclosure of this information, if done inappropriately can bring harm to individuals; this is called a breach of privacy. Getting the PUF design right is essential to a data organization's core dissemination strategy.

Custom Data Files

In addition to, or as an alternative to releasing a PUF, are analyst-prepared ad hoc "custom" data files produced upon special request. While many organizations produce custom files for their stakeholders, NAHDO and the Workgroup do not recommend this approach as an alternative to PUFs, as the "single use" nature of file preparation increases the disclosure risk. Disclosure occurs when information about an individual person represented in the database is released to a party, outside of the data agency with no legal authority for access to the information, which results in potential harm to the individual by revealing personal health information. Custom files should not serve as a replacement for a pre-designed PUF approach to data release for several reasons:

- These "single use" files increase the likelihood of errors and inappropriate disclosures as they are produced in a single run by an analyst in response to a stakeholder request;
- Increased cost and burden associated with responding to individual data requests wastes valuable and shrinking resources.

[1] Public Health Data Dissemination Guidelines: NAHDO working Technical Paper Series available at: http://www.nahdo.org/sites/nahdo.centralpointdev.com/files/Resources/Data_Release_Access_and_Pricing/PH%20Data%20Dissemination%20Guidelines-2005.pdf

Research Data Files

Research data files may retain direct and indirect identifiers as well as potentially sensitive data elements necessary for many research and linkage projects. Therefore, these files are tightly controlled and are usually only provided if approved by the data agency oversight body and/or an authorized Institutional Review Board(s) (IRBs), either the data agency's or the research institution's (and often both). An IRB request for release must address the purpose, funding source, methodologies, security and data access provisions, the risk to the patients, and describe how results will be published. In addition to IRB oversight, research releases require a very specific Data Use Agreement (DUA) to be signed by the data steward and the researcher.

Intra-Agency/Public Health Specific Release

When state health data organizations are sited in locations other than the State Public Health Unit, there may be special provisions in statute or administrative rules which allow release of files with patient direct identifiers to the state and local public health entities. Other state agencies may also receive more detailed information; for example, the approved, qualified individuals in the state Medicaid program or Insurance Commission may have special access to this more detailed data. Generally, these releases are guided by statute, administrative rule, or a data oversight board.

Public Use Files: How They Are Used

A PUF is released to the "public" and it discloses information about the hospitalizations that occurred within the state. Limitations and modifications of specific data elements collected by states assure that what is released will not inappropriately re-identify individual patients. While the amount of information in each data element released in the PUF is limited, the value for end-users is the capacity to aggregate a large volume of data and produce statistical reports providing answers to important questions on hospital care within a state. The Workgroup underscored the fact that state data agencies have had hospital discharge PUFs for many years without a known single disclosure case of patient identity.

PUFs are tested and approved by agency leadership and data boards and are restricted in the amount of information they contain through data release policies and methods, which can vary across states. The PUFs are released in formats without direct identifiers in all states; some states also specify certain other information they collect as indirect identifiers and remove or mask those data elements. For purposes of these guidelines, direct and indirect identifiers are considered as follows:

- A direct patient identifier, "Is information that identifies an individual or can be used to identify the individual and includes: an individual's name, social security number, street address, subscriber number, or other unique data element associated with the individual"? States may receive direct identifiers from hospitals or other health care providers if authorized in statute or rule.
- Indirect patient identifiers include data elements that are not unique to the individual but which contain descriptive information about the person or event, such as: date of birth or dates related to the inpatient stay. While the data organizations maintain direct and indirect identifiers in their original files, new files are prepared where these indirect identifiers are modified (for example, discharge date will become discharge quarter) or removed with substitutions, e.g., date of birth will be removed and in substitution age category will be included. The new file containing modified indirect identifiers assures that an individual could not be identified from the combination of data elements in the file.

The most common use of a PUF is to create aggregations of hospitalizations. For example, aggregating hospitalizations with diagnoses for diabetes allows an end-user to issue a report on the increase in hospitalizations for diabetes in a state, and this could drive changes in healthcare benefits. The information can be further aggregated across states for use by federal agencies. While the PUFs provide a snapshot in time of hospital care, the ongoing collection and maintenance of these PUFs allows for longitudinal analyses of changes over time in hospital care, patient survival, types of procedures done in hospitals, diagnoses, charges for care, and other public health uses, including surveillance of diseases, environmental impacts on health, etc.

The Legal Context for Public Use Files

PUFs are released according to state laws and administrative rules which guide both the healthcare data that is submitted by healthcare providers and the format and items allowed for release of health information to external parties for a variety of uses. While most states collect standard Uniform Bill (UB) data elements (used by healthcare providers and insurers) in their hospital discharge systems, some states do not collect all of the items on the Uniform Bill. Therefore, the statutes and rules vary according to which specific elements (direct and indirect identifiers) can be released.

The HIPAA Privacy Rule's Impact on State Health Data Organizations

The **Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule**[2] establishes regulations for the use and disclosure of Protected Health Information (PHI). HIPAA regulates the use and disclosure of certain information held by "covered entities" (generally, health care clearinghouses, employer sponsored health plans, health insurers, and medical service providers that engage in certain transactions). Most state health data organizations are not considered "covered entities" under the HIPAA privacy provisions of federal law. With over forty states collecting hospital data under legislative mandate, more stringent state laws and administrative rules are likely to govern state health data organizations and their data release policies. In this document we will not address the special provisions required for "covered entities" as they release PUFs.

HIPAA Privacy Rule provisions have heightened state legislators concern about data privacy, and as a result, in some states new protections have moved into statutory language related to state health data organizations, even when those organizations are not considered "covered" under the provisions. This guidance document can provide information to legislators on how to protect privacy, but still release meaningful data.

Data Elements Commonly Released in an Inpatient Public Use File

As part of this project, NAHDO reviewed PUFs from 18 states. Not all states collect exactly the same fields from the Uniform Bill (most states collect fields using the UB)—the list below provides the most common elements released in state PUFs and those that the workgroup suggest should be released as a minimum core. They are listed in 4 categories: 1) Not restricted elements (generally addressing utilization of care and includes facility information); 2) Restricted data elements (certain patient demographics—often the indirect identifiers); 3) Enhanced data elements; and 4) Not released data elements (See Figure 1).

[2] HIPAA available at http://www.gpo.gov/fdsys/pkg/PLAW-104publ191/content-detail.html

<u>Not Restricted Elements (LOW CONTROL)</u>

These data elements can be released without further modification as they are not considered alone or in combination as identifiable of the individual patient. These data elements are useful to health services research, public health, market studies, and other applications. Data elements in this category are those that are included in the PUF without additional modification; they do not contain information that would lead to the re-identification of a patient. In order to provide some of these elements, the originally collected fields have been modified/created/aggregated to reduce risk of re-identification. For example discharge data has been modified to discharge quarter and facility charges may have been rounded to the nearest $100.

Utilization Data Elements

- Type of care/type of service
- Source of admission
- Type of admission
- Length of stay
- Admission quarter
- Admission day of week
- Admission year
- Discharge quarter
- Discharge day of the week
- Discharge status
- Primary payer category
- Secondary payer category
- Total charges
- Facility charges
- Professional charges
- Revenue codes
- External cause of injury-principal E code
- Other E codes (range from 2-10)
- Present on Admission (POA)-principal E code
- Other POA E codes (specify number released)
- Admitting diagnosis
- Principal diagnosis
- POA-principal diagnosis
- Other/secondary diagnosis (range from 3-30)
- POA-other diagnoses

- Major Diagnostic Category (MDC)
- MS-DRG Grouper version
- Principal procedure
- Other/secondary procedures (range from 3-30)
- Principal procedure days
- Agency-assigned record number

Facility Data Elements (*Not restricted for patient protection---only restricted in states with provider protection policies.*)

- Provider Identifier/ Hospital Identification Number
- Facility/Hospital National Provider Identifier (NPI)

Restricted Elements (MODERATE CONTROL)

Elements in this group are typically released in a PUF but are often modified when a combination of these data elements creates a unique record (described in the following section of the report). When unique records occur, agencies modify one or more data elements to assure only non-unique records are contained in the PUF. The practices are described in the following section of the report.

Patient Demographic Elements

- Age in years
- Age groups
- Gender
- Race
- Ethnicity
- State of residence
- County of residence
- Patient zip

Enhanced PUF Elements

These data elements may not directly or indirectly identify patients but are not commonly included in state public use files. Not all states capture these data elements or have the resources to monitor their quality. The Workgroup agreed that if collected states should consider including in their PUFs.

- Birth weight newborn
- Physician specialty
- Physician identifiers (NPI, State License Number (SLN), specialty for attending, operating and other physician)
- Readmission flag/indicator
- Do not resuscitate
- Transfer flag/indicator

Not Released Data Elements (HIGH CONTROL)

These data elements are not disclosed in a PUF.

- Patient name
- Patient address
- Date of birth
- Patient Social Security Number (SSN) or unique identifier, non-encrypted
- Dates (admission, discharge)
- Birth year
- Age in days
- Age in months
- Plan/employer group number

Common De-Identification/Statistical Modification Techniques of Patient Demographic Data Elements

The Workgroup assessed current state PUF practices which have been used by states for decades without known re-identification of individuals. The Workgroup concluded that there is no single approach that will meet every unique need across states and acknowledged that there is no perfect solution. However, they agreed that documentation of common state practices will improve PUF utility and promote consistency across states. Drawing on proven common state practices can help reduce the need for states to create idiosyncratic approaches which increase cost in a time of limited resources and may unnecessarily reduce the utility and reliability of a state's PUF. A poorly conducted de-identification of a PUF may lead to bad science and bad decisions.[3] Thus, NAHDO recommends states follow existing and common practices. This document provides guidance on practices that modify patient demographic and select utilization data elements. Clinical data elements (diagnosis and procedure codes) are extremely useful for end users and should not be modified (unless state law has specific release provisions). The Workgroup agreed that states/agencies should engage their local stakeholders in the design of their PUFs to create a file that has utility and broad support and buy-in as well as compliance with legal provisions.

Philosophical/Judgment Approaches for Statistical Modification

The workgroup spent some time discussing that as a state designs its data protection protocols, along with statistical solutions, an element of judgment is essential. Statistical de-identification seeks to protect and balance two vitally important societal interests: 1) Protection of privacy and 2) Preservation of the utility and accuracy of statistical analyses performed with de-identified data. States must take these factors into account as well as the laws, values, and population distributions in their state. These different factors and perspectives influence how certain data elements might be released. A data element, such as ethnicity, may not increase risk of identification of an individual in a highly populated state such as California, but in states with small populations, the risk increases. Therefore, in these states, additional efforts to create groupings or categorizations have been used to mitigate risk.

[3] "On k-Anonymity and the Curse of Dimensionality" by C. Aggarwal http://www.vldb2005.org/program/paper/fri/p901-aggarwal.pdf

Some states, with heightened data protection provisions, may sacrifice important information by removing what they define as “sensitive” data elements. For example, some state laws prohibit the release of a HIV diagnosis or Mental Health diagnosis which they consider to be a release of ‘sensitive information’ and may remove these diagnosis codes from the files. NAHDO and the Workgroup consider removal of these important elements such as diagnoses, problematic and would instead recommend modification of demographic elements or up-grouping diagnosis codes to avoid sensitive information release of individuals while maintaining important information. For example, removing all mental health diagnoses (primary and secondary) will impact the validity of the file, given that about one-third of all adults receiving hospital care have a co-morbidity related to mental illness, requiring filtering of a significant number of records. States must balance utility of the file and varying methods of privacy protection. Many end users of the files utilize risk adjustment tools to level the playing field for hospital comparisons; hospitals providing care to “sicker patients” benefit from robust risk adjustment models including co-morbidities, such as mental health conditions. Fairness requires complete information, especially when provider performance information is publicly reported or when engaged in pay-for-performance programs. Understanding how the data is used is critical when making decisions about removal versus modification of elements and requires an element of judgment.

Defining a Unique Record in the Data

The Workgroup spent some time discussing the need to be clear about why data is modified for a PUF and which modification technique is appropriate. The Workgroup agreed that the reason a PUF is modified is to be sure there are no unique demographic characteristics retained in the file, so first it’s important to define the boundaries/parameters for a unique record. Next it’s important to decide how to identify unique records in the PUF, and finally there is a need to decide which technique(s) to apply to eliminate unique records from the PUF.

First, a state should define what is meant by the term unique record in the available data set. As already mentioned the Workgroup agreed, unless specified by law, clinical data elements (diagnosis and procedure codes) should not be modified, and therefore demographic data elements should be those considered when determining if a record is unique. When looking at what defines a unique record and the way to apply the methodological techniques it’s important to decide if this is based on:

1. Single demographic fields which make a record unique, or a
2. Combination of demographic fields that make a record unique

States use various thresholds for determining uniqueness of records in a PUF. Thresholds range from two to thirty records with a common practice identified as five. For example if there are five or less records based on a single field or a combination of fields (determined by your state) that are unique then these are typically masked (Note: These five records may be exactly the same, that is, all fields are identical—but still too unique for the file in that the potential risk of re-identification is high).

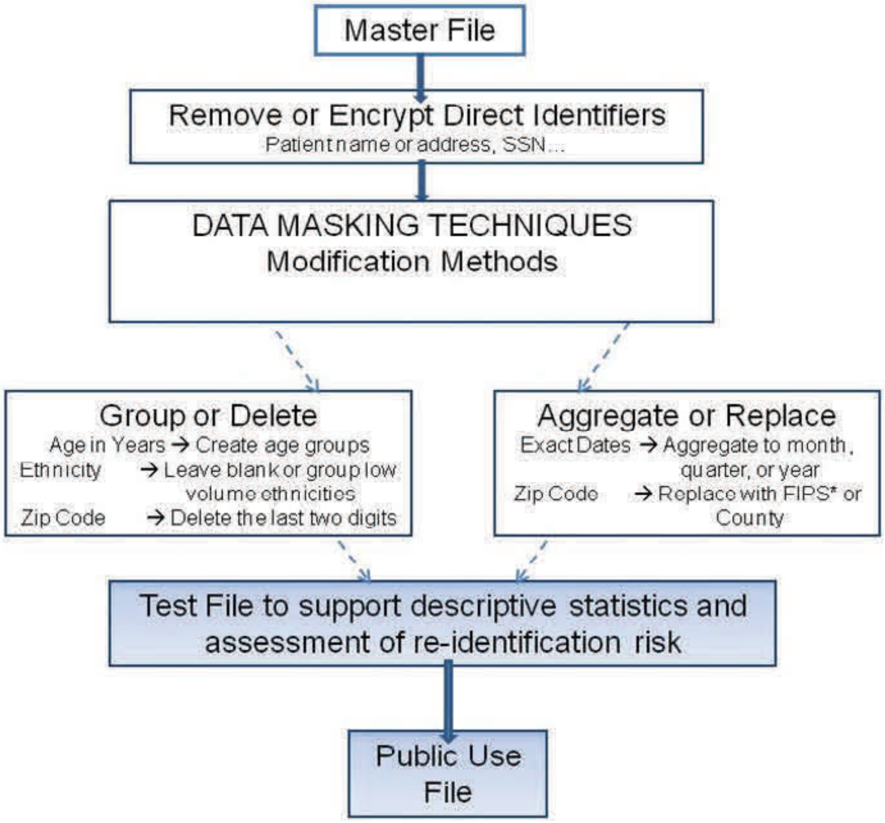
Statistical Approaches for Modification

The common statistical practice identified is to avoid providing information about a small number of individuals whose characteristics/demographics are the same. It is thought that with enough supplemental information the individual might be able to be re-identified, but modification makes it more difficult to do so.

The Workgroup reviewed and discussed a range of practices that states are using to modify the fields in the PUFs to be sure there are no unique records. Below is a description of several statistical approaches/methodologies the Workgroup identified and agreed on as a useful pathway for creating a

PUF. Figure 2 below begins with the identifiable detailed master file that is retained in a secure computing environment with access limited to qualified, authorized individuals. When preparing products for data release and analytic files, the data programmer will first remove direct patient identifiers, such as patient name and address from all records and recode indirect patient identifiers such as patient demographic fields. Preparation of the file from this point on is directed towards the creation of a micro-level PUF. The steps are described in the text below.

Figure 2: Modification Methods for PUF Creation


Master File
Remove or Encrypt Direct Identifiers
Patient name or address, SSN…
DATA MASKING TECHNIQUES
Modification Methods
Group or Delete
Age in Years → Create age groups
Ethnicity → Leave blank or group low volume ethnicities
Zip Code → Delete the last two digits
Aggregate or Replace
Exact Dates → Aggregate to month, quarter, or year
Zip Code → Replace with FIPS* or County
Test File to support descriptive statistics and assessment of re-identification risk
Public Use File

Mapping Zip to FIPS* (Federal Information Standards Code)
State FIPS: code given to each US territory. This code simply identifies each US state or territory by its unique 2 digit code
County FIPS: given to counties or parishes. This code simply identifies each US county by its unique 3 digit code
Software available at: http://www.zip-codes.com/zip-code-database.asp

Data Masking Techniques

Data masking techniques include recoding or restriction of demographic detail that increase the population size or the level of geographic information. Agencies use various types of masking methods which are applied to indirect patient identifiers. Masking methods include recoding of data such as grouping, deleting, aggregating, and replacing data elements. The level of recoding requires a detailed knowledge of the underlying master file and user information needs, as well as an element of judgment provided by the data oversight body. The approach is determined by the laws and policies of each state or data organization and the extent of the data management and other controls in place, such as data use agreements discussed in the final section of this document.

Group or Delete

Grouping of data elements involves splitting a single data element, like age in years, into a larger pool of ages, such as a five-year age grouping to avoid retention of unique information about the patient. Deletion methods modify an existing data element by removing a portion of the information, such as the last two digits of a zip code (Note: each consecutive digit of a zip code provides intelligence). This is a better alternative than complete removal of the data element, as it retains some useful information while reducing re-identification risk.

Aggregate or Replace

Aggregation involves the summing or combining of specific data elements into a single, broader pool of information, such as an admission date to a quarter or year. Replacement is substituting an alternative data element, such as the Federal Information Postal Standard (FIPS) or county or state for the patient zip code. This replacement increases the population size.

California Example
CA removes all the direct identifiers. When determining which demographic variables would be modified (masked), they asked the data users which were the most useful, that is which elements to mask first and which to leave unmasked if possible. Following is the list of elements they selected ranked in order from mask first to mask last.

Age in Years (at admission)
Ethnicity
Race
Sex
Age Range (20 categories)
Age Range (5 categories)
Admission Quarter
Patient ZIP Code (5-digit)
Small County Groups
Patient ZIP Code (3-digit)

The CA process is to take the file (stripped of direct identifiers) and apply a computer program which identifies all unique records in the file. If unique records are found, they begin to de-identify them by masking age in years, run the same program and see which records are still unique. Of those that are still unique they mask ethnicity and so on, right down to the three digit zip. This technique assures that no individual is uniquely identifiable. CAs policy is to include all clinical elements (procedure codes and diagnosis codes etc.) in the PUF; no masking procedures are applied to these clinical elements. One challenge with this method is that as more elements are added to the file, such as preferred language and lab results, they have to make a decision on how to include (mask or not) or whether to include those elements in the PUF.

Example of alternative approaches to data masking:

We share an example where either a grouping or deletion approach would result in the same objective of masking patient identity. For diagnoses (like HIV/AIDS), an agency may delete the last digits of the zip code or recode up to a county or state level where the denominator for this characteristic is larger. Alternatively, an agency could code HIV to a higher level diagnostic category such as the Diagnosis-Related Grouping (DRG). The first method better allows public health users to track the prevalence of HIV disease within the state. The second approach groups HIV with other infectious conditions, losing some fidelity in the process. Other state practices are included in Appendix 1.

Additional Testing Approaches

After creating the PUF, having eliminated all direct identifiers, and modifying as needed the indirect identifiers, data organizations can take some additional steps to test whether the algorithms used on the PUF will assure that individuals will not be re-identified. Internal reports are generated to identify records with unique demographic characteristics (e.g., a single case of HIV in a zip code). Many states have developed SAS programs for their internal testing and have indicated they would share the source code with other state health data organizations. Some states use vendor solutions, an example of which is included in Appendix 2.

Note: Most states provide PUF documentation in a user's guide, and describe the methods used to mask information, as well as definitions, time frames, limitations, caveats, and other technical instructions. (Links included at the end of the document)

Management and Institutional Controls for the Protection of a PUF

States utilize multiple strategies that promote data protection and facilitate data access. Aggregate tables and reports require little to no controls, as the re-identification risk is minimal. Because they are record level data sets, controlling the disclosure or release risk in hospital PUFs is best with a combination of approaches that together provide layers of protection. This section will discuss the additive protection that management and institutional controls provide to protecting the data released in PUFs.

The Workgroup acknowledged that the release of data in a PUF represents a series of trade-offs. These trade-offs include statistical and management controls deployed to reduce disclosure risk, while preserving as much data utility as feasible. More statistical restrictions diminish the utility of the data set while more restrictive management/institutional controls may permit the retention of more granular and useful data. Thus, it is important to note that agencies that apply multiple controls are more likely to release more detailed data. Statistical and management controls deployed include:

1. Statistical/technical modifications that alter the data to reduce the probability that individuals can be uniquely identified/re-identified in the PUF. (These are covered in the previous section of this document).
2. Institutional policies that provide database oversight and support the education of data users. Oversight in the form of boards or advisory committees to design release policies and govern release promotes consistency in release practices. Training and education of users about appropriate use and handling of PUFs also reduces inadvertent re-identification risk.
3. Regulatory strategies to manage and oversee user access to the PUF. These include the use of data use agreements that clarify the terms of PUF use and spell out the legal and financial penalties for inappropriate disclosure of PUF and/or sanctions that prohibit the user from acquiring a PUF in the future.
4. Information Technology (IT) protections for data transfer/distribution that include encryption of media and/or secure file transfer protocols.
5. Transparency mechanisms that provide notice of data requests and the decisions (to release or not) on a state data agency website.

> *“The Maine Health Data Organization notifies the public of all data requests by posting to their website for a ten-day comment period. For requests including identifiable practitioner data elements or the insured group or policy number, interested parties have thirty days to comment. Within the thirty days, the MHDO establishes an advisory committee to assist in reviewing the data request composed of members of their Board, the Maine Health Data Processing Board, the Maine Quality Forum, and representatives from the Maine-based practitioner professional or affiliated medical specialty organizations associated with the potentially impacted practitioners.”*

Institutional Policies and Practices

Data Governance Board

Many data agencies have some form of data oversight board or body that establishes and monitors the policies that govern data release. In some cases, the data board is established by law, but in other cases the agency will appoint an advisory board or utilize an existing agency committee. A data oversight body should be comprised of all key stakeholder groups and be transparent in how it is governed. This oversight is a very important component of data release, for the following reasons:

- Development of comprehensive data release policies that are consistent with state laws, agency policies, and community practices.
- Consistency in how the PUF and related products are released over time irrespective of agency staff turnover.
- Multi-stakeholder participation assists in ensuring fairness of release policies across all users.
- Legal protection for the agency.

Examples of State Data Boards

- Utah Health Data Committee, statutory committee
- Pennsylvania Healthcare Cost Containment Council(PHC4)
- South Carolina Data Oversight Board

Education/Training

In addition to data release oversight, agencies will employ additional measures to assure internal and external compliance to legal and policy restrictions. Ideally, the data-providing agency educates users about the disclosure risk of micro-data, the types of analyses that are considered breaches of confidentiality, and the legal issues associated with disclosure. This education can be in the form of seminars, user groups, or webinars and may be especially important in the early years of PUF distribution or in agencies with global intra-agency use agreements.

Increasingly, public health agencies are using hospital discharge data sets for surveillance, program evaluation, and community assessment initiatives, opening up a huge user base for the data. These public programs may be covered by a single data use agreement covering multiple users.

Release of one data file to the umbrella public health department for use by qualified agencies and personnel within that department under one agreement reduces administrative burden on the data agency and encourages use of the data for the public good. Mechanisms for educating the workforce and monitoring compliance should be considered.

Both the Washington State Department of Health and the California Office for Statewide Health Planning and Development have held Data Users Conferences. In-person training requires a substantial investment by the data-providing agency, which in today's budget environment may not be possible. Less costly methods can be implemented via web seminars. Regardless of training approach, incorporating user responsibility and constraints directly into the data use agreement or data licensing agreements, requiring the user to attest that s/he will comply with the terms and that s/he is aware of the penalties for non-compliance, complements and reinforces the provisions of use.

Regulatory Approaches

Data Use or Data Protection Agreements (DUA or DPA)

Data use agreements (DUA), also called data protection agreements (DPA), are agreements that must be executed, prior to the disclosure of data (in this case, PUF) by the data steward agency, to ensure that the disclosure will comply with the requirements of all relevant laws and agency data release policies. While direct patient identifiers are not contained within a PUF, there may be indirect identifiers, which alone or in combination with other publicly available information, could lead to the identification of individuals, should the data recipient work to re-identify the data. Hence, many states require that the agreement must be completed prior to the release of, or access to, the PUF.

Since PUFs do not contain any individual direct patient identifiers, some data agencies may not require them. However, a DUA containing restrictions promotes user self-regulation and penalties/sanctions for non-compliance and can serve as an educational tool for users. Most states request users of the PUF sign a DUA (and may require legal notarization) before gaining access to the data.

Summary of Common Clauses in State DUAs

A solid and clear DUA/DPA that is more restrictive in how the user can use and release the PUF will usually permit the agency to include more robust data in their PUF. A DUA/DPA will usually contain the following provisions:

- Data within the PUF or the PUF itself cannot be re-transferred or re-disseminated in a format that could possibly lead to identification of an individual.
- Prohibits any attempt to re-identify (individual patient identity) from records in the dataset. When using healthcare data, a data user may identify an individual in the database using knowledge gained from other sources (including personal knowledge of the hospitalization) even if the data user had no intention of identification. Therefore, this component of the DUA adds a layer of protection by prohibiting this type of inadvertent attempt at re-identification. Prohibits attempts to identify or contact subjects represented in the data set.
- Prohibits the disclosure or sharing of the data in ways other than stated in the agreement, or as otherwise required by law. Most agency DUAs specify who is permitted to access and use the data and prohibit patient identification in publications/dissemination. Typically this means using the data only for the purpose outlined in the application and for no other purpose, with the DUA having a statement along the lines of "I will use the data only for the purpose identified in my application…"

- Most states prohibit linkage using the data contained within the PUF; For example, "I will not attempt to link, nor permit others to attempt to link, any record from the data file with individually-identifiable data from any other source." The Workgroup made a distinction between types and intent of linkage. Patient-level linkage is prohibited, but facility or area level linkages are generally permissible. The DUA/DPA should consider the intent of the linkage, with re-identification of patients as prohibited. Some linkage may be allowed for public health uses under their data agreement, statutes or administrative rules.
- The user is responsible (not the agency) for analysis of the data—and must agree to use data responsibly in their analysis. When publishing results of their analysis, authors (users) are asked to provide a notification that indicates the data agency is not responsible for the analysis—they should indicate the analysis was done by the author and that interpretation is from the user not the state.
- Requests that the data source/agency is cited in publications.
- Requires the reporting of any misuse or unauthorized disclosure as soon as known.
- Ensures that any agents, including subcontractors, agree and are bound to the restrictions and conditions of the DUA. All agents should be named in the agreement.
- Includes a penalty clause for non-compliance with the DUA terms; these penalties range from revoking access to data to the imposition of fines and/or imprisonment as defined in state law or regulations.

States vary in the style/format of their DUAs for the release of PUF, ranging from a single page assertion to detailed applications where the data recipients are required to spell out the specifics of how they will use the data and the assurance that security and confidentiality will be protected. Both the data steward (person releasing) and recipient must sign the agreement. These assurances and signatures are often contained within the data request form (Links to DUAs are listed at the end of the report).

Information Technology Protections

Information Technology strategies include encryption for CD-ROMs and DVDs containing PUFs and transmission by secure FTP for electronic files. An added layer of protection for PUFs may include user authentication or password protection for various media like CD-ROMs and DVDs containing PUF content (if being used with zip program to encrypt at the same time—no additional software required). This is relatively inexpensive and is not technically difficult to implement. This technique provides added protection and supports tracking of users. Administration costs include set-up and maintenance of a logon/identification process, and monitoring of use. This is a relatively low-cost method of adding an extra layer of security to PUFs and does not alter the underlying data; however, there are costs associated with assigning and administering the security measures.

Other Regulatory Controls

PUF pricing, while driven by many factors including demand, state law, and agency funding mix, is an indirect strategy for controlling access to micro-level data. By pricing a PUF at a higher level, it may deter the most casual or occasional user. While pricing alone is far from a data protection strategy, it may serve, in conjunction with other data modification methods and regulatory controls, to limit the access and guide the use of detailed micro-data. The trade-off is that higher prices may limit legitimate user access, so this should be taken into the pricing consideration.

Other regulatory controls include licensing of the PUF and pricing strategies that differentiate between the single individual use or user and a corporate user that will repackage and redistribute a subset of the data elements in proprietary reports and data products.

In some states, agencies use the public health authority to extend data access and use; this permits broad access to discharge data across state public health programs, and streamlines the data application process. Using this approach effectively requires the inclusion of education and training of the public health workforce by the health data organization as to the legal, technical, and political considerations. Under these circumstances we suggest that:

- The DUA could serve as intra agency agreement---defining how the file is shared across different programs for public health purposes;
- The DUA could also require the provision of training within the agency on use of the data.

Conclusions

Health data agencies that maintain large-scale health care data bases have unique missions to disseminate information and have implemented efficient and effective dissemination strategies, including the release of de-identified data files for multiple users and uses. With this guidance, data agencies releasing PUFs have a tool to review and compare their current release practices with those of other states. NAHDO recommends that states use multiple data modification approaches that address the individual characteristics of their state's demographics and couple these modifications with management and institutional controls to govern the release and use of a PUF. These approaches are based on existing and proven "best" practices that states have successfully employed for many years. Because a "one-size-fits-all" solution rarely fits the needs of every state, we have documented examples of common approaches. Each method for modifying data and controlling its use involves trade-offs to balance the utility and broad use of information with the protection of the privacy of the underlying data. While we work towards more uniform methods for assuring that individuals are not at risk of identification, we must also realize the uniqueness of each state and not prescribe a single approach to de-identification.

This guidance document will evolve as state agencies respond to new technologies and applications.

NAHDO maintains health data organization profiles on its website that includes links to state legislation, technical specifications, and data release information (including DUAs). This information is provided for no charge to NAHDO members and for a fee to others. This can be found at www.nahdo.org

Appendix 1: Example State Practices

The following table describes some of the modification techniques/practices states have adopted for specific or set of data elements as well as the Workgroup's recommendation.

Data Elements	Example Practices Identified	Notes/Considerations
Usually not restricted elements		
Facility Data Elements Facility/Provider Identifier/Hospital Identification Number Facility/Hospital National Provider Identifier (NPI)	State Data Organizations assign facility identifiers and most include these in a PUF.	Unless prohibited by law, states should release facility/hospital-specific information. This information adds utility to the public data set and, with few exceptions, cannot directly or indirectly identify individual patients.
Typically modified elements----Patient Demographics in PUF: These data elements alone will not directly or indirectly identify individual patients but by combining two or more of these data elements the size of the population of interest is reduced, making it easier to identify individuals within the data base.		
Age	States typically map Date of Birth into age grouping (age in years or age categories). Example, for ages 5 years to 84, group patient age in five-year cohorts at admission. Less than 5 years of age: Under one year category 1-4 for children 1 to 4 years of age Over 84 years: >85	These groupings vary by state and should be designed to be relevant to the state and the state's users.
Gender	Limit breakouts to male/female	Some states capture multiple gender categories; states may group unknown/missing/other together for release, if relevant to that state.
Race and Ethnicity (R/E)	Limit breakouts to major categories such as black, white, Asian or other---not as minor race categories such as Sudanese, Thai etc..	If R/E are included in PUF, release according to cell-size rules for state. (Larger states are more likely to include R/E on PUF, smaller states may not release on PUF; only on research or custom release)

Patient County	The majority of states include county as a variable and limit/ modify release of demographics if cell sizes are too small. For unique records, some states release geography only at state level instead of modifying demographics.	For counties with small populations, there is more risk of re-identification for one or more unique records and states may consider suppressing.
Patient Zip	The majority of states include zip code as a variable and limit/ modify release of demographics if cell sizes are too small **Practices Identified in States:** Map zip to FIPS Follow state cell-size rule; drop or encrypt last three digits, or aggregate into a larger geo area For unique records, release geography only at state level instead of modifying demographics.	This data element adds utility to the public use file and states should select the most appropriate approach.
Utilization Information---these elements not modified		
Examples of these elements: **Type of care/type of service** **Source of admission** **Type of admission** **Admission quarter** **Discharge quarter** **Discharge status** **Primary payer category** **Total charges** **Facility charges** **External Cause of Injury-Principal Ecode** **Admitting diagnosis** **Principal diagnosis** **POA-principal diagnosis** **MS-DRG Grouper version**	Total and Facility Charges: Round total and facility charges to the nearest $100 to avoid linkage with other databases to identify an individual patient.	Diagnosis/Ecodes: The rarer the condition, the easier it is to identify an individual in the database, since very little additional information is needed to uniquely identify a patient. Some states have legal prohibitions to the release of specific conditions (HIV/AIDS, Mental Health, Substance Abuse).

Sometimes restricted		
Physician identifiers (NPI, SLN, Specialty for attending, Operating and other physician)	Not all states release at the physician level.	The fields themselves may not lead to identification of a patient, but in combination with other fields.
Enhanced File Data Elements: These data elements may not directly or indirectly identify patients but are not commonly included in state public use files. Not all states capture these data elements.		
Birth weight of Newborn	If a state releases birth weight on a public use file, rounding up to the nearest 500 grams to avoid linkage with other data to identify an individual patient. (This information can also be obtained through the ICD-9 code category)	While birth weight of newborns is useful for understanding changing obstetric practice, only a few states release on public use files.

Appendix 2: Vendor Solution: Using a Tool to Measure Disclosure Risk and Apply Protections.

A commercial tool can be helpful in creating a PUF that balances the competing goals of maximizing data released and minimizing disclosure risk. A tool should be able to ingest micro-data records, identify risk, and provide options for mitigating that risk in the data to be released.

OptTek Systems has created a tool called OptShield™ that can be used to measure the risk of potential disclosure in a PUF and to automate the removal or aggregation of identifying data elements to attain an acceptable level of disclosure risk.

The flow chart below illustrates the main steps in the PUF creation process and highlights how OptShield is used. Once historical data has been collected it can be loaded into OptShield from a comma delimited file where each row represents a single record. Next the user specifies which elements in a data record constitute potentially identifying data. The identifying data generally includes all patient demographic elements labeled as Restricted Elements above. However, the user of the tool determines at this step which elements to treat as potentially identifying and can specify a larger or smaller set of elements.

Next, the tool measures disclosure risk for the data set. The tool supports multiple measurements of risk. The most common measurement is determined by assembling all individual records into groups that share the exact values for all identifying elements. If there are groups with small numbers of records these are considered disclosure risks. The user of the tool must determine what the acceptable minimum number of records in a group should be. If the disclosure risk is acceptable, i.e. the smallest group of records contains more records than the minimum acceptable number of records, then the PUF can be published as is. If not, the user can use the tool to either manually identify specific identifying elements to remove or aggregate, or the user can have the tool automatically determine the best set of identifying elements to remove or aggregate.

The path on the left side of the flow chart represents the tool assisted manual approach. The user will use the tool to specify an identifying element to remove or to aggregate. The tool will remove or aggregate the element and then measure disclosure risk for the adjusted data set again. If the acceptable disclosure risk level has been met, then the PUF can be published. If not, the user repeats the manual process.

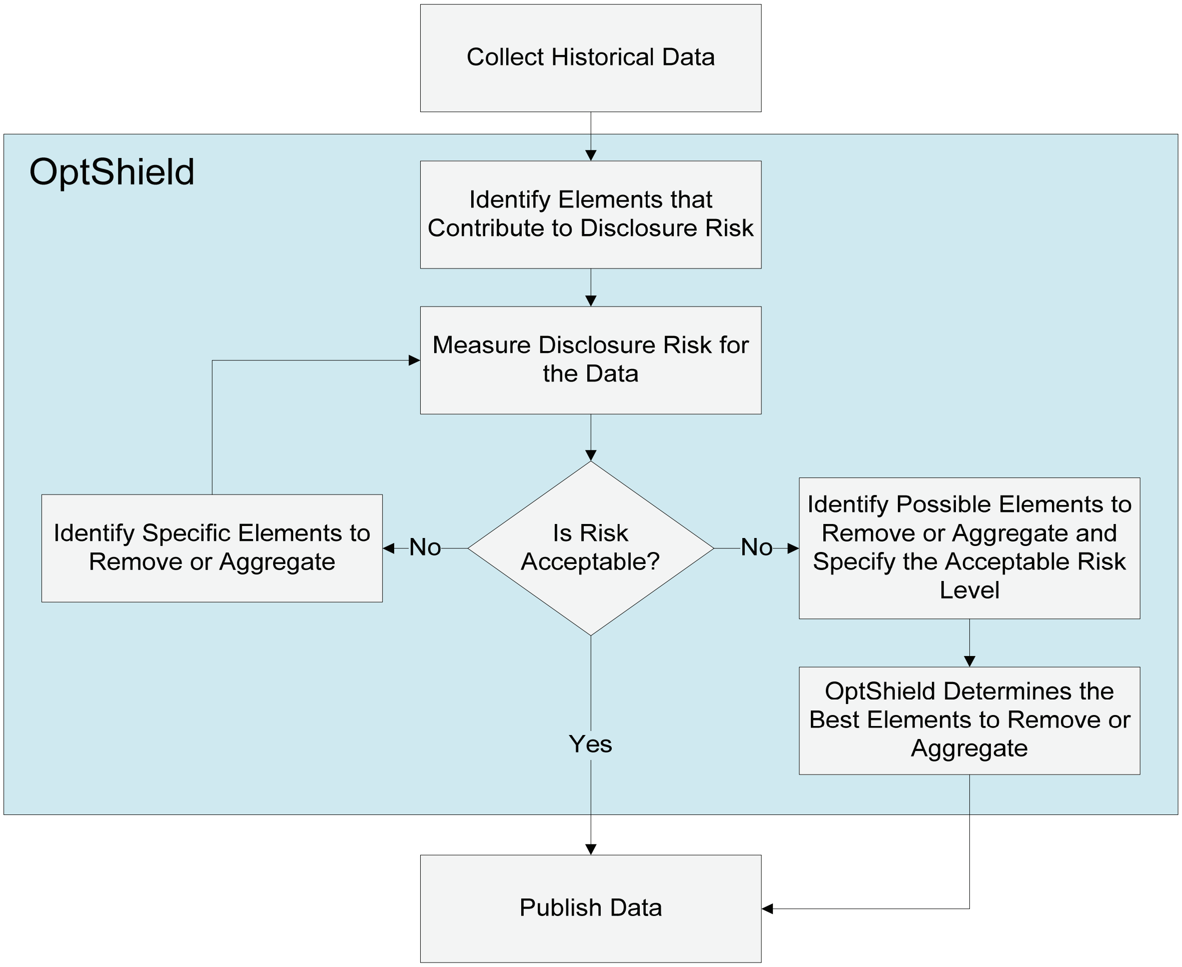
The path on the right side of the flow chart represents an automated option provided by OptShield for determining the best set of identifying elements to remove or aggregate. There are two ways that "best" is measured: (1) By minimizing the simple sum or a weighted sum of the number of identifying elements removed or aggregated; and (2) By minimizing the difference between the following:

- the number of unique groups created originally by collecting all individual records into groups that share the exact values for all identifying elements
- the number of unique groups in the final adjusted data set by collecting all individual records into groups that share the exact values for all identifying elements

If the user wants OptShield to automatically generate a protected data set, they specify the following:

1. The way to measure "best"
2. The possible elements to remove or aggregate
3. The acceptable risk level

OptShield then searches for, and finds, the best set of identifying elements to remove or aggregate to achieve the best protected data set.


Collect Historical Data
OptShield
Identify Elements that Contribute to Disclosure Risk
Measure Disclosure Risk for the Data
Is Risk Acceptable?
No
Identify Specific Elements to Remove or Aggregate
No
Identify Possible Elements to Remove or Aggregate and Specify the Acceptable Risk Level
OptShield Determines the Best Elements to Remove or Aggregate
Yes
Publish Data

Consider a small example with twenty records where each record contains Hospital, Condition Code, Gender, Age, and Zip Code.

Hospital	Condition Code	Gender	Age	Zip Code
Plains Medical Center	3	F	64	78701
Community Hospital	5	F	72	78701
Mercy Medical Center	6	M	25	78701
Plains Medical Center	1	M	35	78701
Community Hospital	4	M	31	78701

Memorial Hospital	7	M	54	78701
Mercy Medical Center	6	M	78	78701
Community Hospital	2	F	36	78702
Plains Medical Center	6	F	48	78702
Community Hospital	7	F	41	78702
Memorial Hospital	4	F	68	78702
Memorial Hospital	2	M	22	78702
Mercy Medical Center	1	M	59	78702
Plains Medical Center	4	M	60	78702
Mercy Medical Center	4	F	69	78703
Memorial Hospital	6	F	79	78703
Community Hospital	6	M	49	78703
Memorial Hospital	5	M	45	78703
Plains Medical Center	6	M	57	78703
Mercy Medical Center	3	M	57	78703

A user can use the OptShield tool to evaluate risk and determine how to remove or aggregate data elements before releasing this data. Assume the user decides to treat Gender, Age, and Zip Code as identifying elements for the data set. The user also determines that after grouping data records by these identifying elements, no group can have less than two rows. Under these conditions all data records are initially unique, i.e. grouping records by Gender, Age, and Zip Code yields 18 groups of one record each, so the user needs to remove or aggregate data elements before releasing this data. Assume the user wants to investigate combinations of one or more of the following options:

- aggregating age into 10 or 20 year age bands, or removing age
- releasing only the first three digits of the zip code
- removing gender

The table below illustrates a set of options for removing or aggregating the identifying data elements. For each option the number of identifiable records (data records which individually form a unique group) and the number of unique groups is shown.

	Identifiable Records	Number of Unique Groups
Release Raw Data	18	19
10 Year Age Bands	12	16
20 Year Age Bands	7	12
Remove Age	0	6
Remove Gender, 10 Year Age Bands	8	14
Remove Gender, 20 Year Age Bands	1	8
3 Digit Zip Code, 10 Year Age Bands	3	10
3 Digit Zip Code, 20 Year Age Bands	1	6
Remove Age, Remove Gender	0	3

In this example there are only two options that provide the level of protection desired for the dataset, "Remove Age" and "Remove Age, Remove Gender". All other options leave at least one identifiable data record in the dataset. Of the two viable options the first, "Remove Age", is superior since it leaves more data in the dataset. This is obvious in this simple example, but in a more realistic example with many options providing adequate protection, the number of unique groups remaining for an option is a good measurement of the amount of useful data remaining in the data set.

For realistic data sets containing from thousands to millions of records, OptShield provides the ability to quickly measure risk and evaluate tradeoffs between aggregating and/or removing identifying data elements.