

Limited or de-identified data sets (Public Use Files)

State health data organizations have been releasing hospital discharge data sets with a reduced number of data elements for public users for many years. In fact, the states' practices served as a model for the HIPAA Privacy Regulations' "Limited Data Set" (LDS) which governs the data release of 'covered entities' subject to the federal HIPAA provisions. The LDS (sometimes referred to as the public use file or PUF) under HIPAA defines 18 data elements that directly identify a patient and which must be removed for the information to be considered de-identified and not protected health information (PHI):

- names;
- all geographic subdivisions smaller than a state, including county, city, street address, precinct, zip code,¹ and their equivalent geocodes;
- all elements of dates (except year) directly related to an individual; all ages >89 and all elements of dates (including year) indicative of such age (except for an aggregate into a single category of age >90);
- telephone numbers;
- fax numbers;
- electronic mail addresses;
- Social Security numbers;
- medical record numbers;
- health plan beneficiary numbers;
- account numbers;
- certificate and license numbers;
- vehicle identifiers and serial numbers, including license plate numbers;
- medical device identifiers and serial numbers;
- Internet universal resource locators (URLs);
- Internet protocol (IP) addresses;
- biometric identifiers, including fingerprints and voice prints;
- full-face photographic images and any comparable images; and

¹ Zip codes are often included in state hospital discharge systems; most states prescribe some limitations related to number of discharges per zip code for release of data.

- any other unique identifying number, characteristic, or code, except that covered identities may, under certain circumstances, assign a code or other means of record identification that allows de-identified information to be re-identified.

State public use file design (de-identification) practices precede HIPAA and utilize approaches that retain the geographic granularity so essential to state market and policy applications. Because many state health data organizations are exempt from the federal provisions under HIPAA, if their state laws governing data collection and release are more restrictive, health data organizations have crafted local release policies that omit direct patient identifiers from public use files, while retaining essential data elements for market and policy studies. These policies are carefully crafted in each state and reflect local decisions about the trade-offs between data utility and patient confidentiality.

Strengths of a limited data set are many. It limits the amount of identifiable data released to users, as a de-identified data set meets the needs for most common applications. The de-identified data set streamlines the data review/release process by not requiring Institutional Review Board (IRB) approval and yet the data set still supports most statistical studies.

The trade-off is that a limited data set results in the loss of specificity in some data elements and the elimination of confidential data elements reduces the capacity for linking data across time, maintaining longitudinal studies, or episode of care analyses. Ironically, these are some of the uses most desired by healthcare providers and policy makers.

Data made available in the state public use files is available in the state data agency profiles.