UCDAVIS HEALTH

DCOE Data Center of Excellence

De-identification of Data for Research Projects

This tip sheet is meant to clarify expectations between researchers and the Data Access and Delivery Team when provisioning data extracts for research. The Data Access and Delivery Team is performing the role of an Honest Broker.

Definitions

There are three ways to remove and/or obfuscate data for research projects that is supported by the Data Access and Delivery Team. They are defined below in the order of the most protected data to less protected but potentially more useful.

1. Anonymized data is a data set that has PHI removed and/or obfuscated to ensure patients cannot be re-identified. This type of data is meant as a one-time extraction where the patients cannot be re-identified even by the Honest Broker as a mapping key is not maintained.
2. De-identified data is a data set similar to anonymized data sets, however, a crosswalk table of the de-identified patient identifier and the original patient identifier from the data source is maintained so that the patients can be re-identified in the even that additional data can be extracted for those patients.
3. Limited Data Sets (LDS) have most PHI removed or obfuscated as in a de-identified data set, but can include limited PHI; namely geographical sublocations and dates.

Reviews and Approvals

- The Data Access and Delivery Team does not require an IRB protocol for de-identified or anonymized data sets, but it does require the protocol and a copy of the UCDH IRB determination letter for Limited Data Sets.
- Limited data sets still require an accounting of disclosures. For data provided to you by the Data Access and Delivery Team within the Data Center of Excellence, this will be done on your behalf.

Data Types

DATA ELEMENT	LIMITED DATA SET	ANONYMIZED AND DE-IDENTIFIED DATA SETS
Names of patients, relatives, employers or household members (Does not include names of providers).	Removed	Removed
Address, city, and other geographic information smaller than state.	Street addresses are removed. Cities, towns, states, and full zip codes can remain.	– Street addresses are removed. – Zip codes are truncated to only show the first three digits. – Zip codes, cities, towns and counties with < 20,000 residents are obfuscated. The zip codes will display as ‘000’. Cities, towns, and counties will display ‘Other/Unknown’.
All elements of dates (except year); plus, age and any date (including	Original dates are retained.	Dates must be obfuscated using one of these methods:

DATA ELEMENT	LIMITED DATA SET	ANONYMIZED AND DE-IDENTIFIED DATA SETS
year) if age is over 89. Examples: date of birth, date of death, date of admission, date of discharge, date of service.	For DOB, only the year is provisioned.	– Display year only – Shift dates by random offsets (same offset per patient) – Share time intervals between events – Age and date of birth are obfuscated. When patient is 90 years old or greater, display patient age as 90.
Telephone, fax numbers; e-mail addresses, web URL addresses, IP addresses.	Removed	Removed
Social security number, medical record number, health plan beneficiary number, any account number, certificate, or license number.	Removed	Removed
Vehicle identifiers and serial numbers, including license plate numbers, Device identifiers and serial numbers, biometric identifiers, indefinable photography.	Removed	Removed
Any other unique identifying number, characteristic or code.	Removed	Removed
Notes	Removed	Removed
Images	Removed	Removed
	DADT is working on a process to share some images de-identified by approved software tools for future use.	

Examples

The following examples may help to illustrate what kind of data would be delivered in a de-identified or anonymized data set.

Zip Codes

Zip Code	Population	What Will Be Shown in the Data Set: Zip Code
00601	18570	006
00602	41520	006
00603	54689	006
55616	6994	000
69201	4005	000
69210	2484	000

Counties

County Name	Population	What Will Be Shown in the Data Set: County
Barbour County	27457	Barbour County
Bibb County	22915	Bibb County
Blount County	57322	Blount County
Bullock County	10914	Other/Unknown

Date Obfuscation: Year Information Only

Event	Date	What Will Be Shown in the Data Set: Year
Enrollment	03/01/2013	2013
First Visit	03/20/2014	2014

Shifting Date by Random Offsets

For each patient, a random integer between +/- 365, excluding zero, is assigned. All dates in each patient record are shifted by the same amount. For example, for patient John Doe, all dates in his patient record may be shifted by adding 35 days.

Shifting dates by a random offset provides the following benefits:

- Preserves order and duration of events
- Preserves the meaning of variables

Patient	Encounter Date	Enrollment Date	Random Integer Assigned to the Patient	What Will Be Shown in the Data Set	
				Encounter Date	Enrollment Date
1	08/05/2020	10/10/2020	22	08/27/2020	11/01/2020
2	05/06/2019	07/08/2019	-50	03/17/2019	05/19/2019
3	09/14/2021	11/01/2021	261	06/02/2022	07/20/2022
4	07/04/2018	09/15/2018	-6	06/28/2018	09/09/2018
5	11/26/2020	01/25/2021	31	12/27/2020	02/25/2021

Time Intervals

To use time intervals dates of events of interest will be shared as a numeric value that represents the difference between the baseline event and the date of the event of interest. The granularity (years, months, weeks, days, hours, etc.) of the interval can be determined by the data requestor.

Below is an example of how many days elapsed between an enrollment and an encounter. The enrollment date would be used as the baseline.

Patient	Enrollment Date	Encounter Date	What Will Be Shown in the Data Set: Days
1	08/05/2020	10/10/2020	66
2	05/06/2019	07/08/2019	63
3	09/14/2021	11/01/2021	48
4	07/04/2018	09/15/2018	73
5	11/26/2020	01/25/2021	60

Age

Ages can be provided in de-identified and anonymized data sets with one exception: if a patient is 90 years old or more, they must be grouped into a single category.

Age	What Will Be Shown in the Data Set: Age
12	12
34	34
89	89
90	90
96	90

Year of birth

When the difference between the year of birth and current date results in the patient being over 90, the year of birth needs to be adjusted so the patient's age will reflect 90 years.

Date of Birth	Current Year	Age	What Will Be Shown in the Data Set: Year of Birth
01/01/2010	2022	12	2010
01/01/1981	2022	41	1981
01/01/1933	2022	89	1933
01/01/1932	2022	90	1932
01/01/1928	2022	94	1932