

**UC Tech Awards 2023 Form**

**Category: IT SECURITY**
**Name:** Center of Excellence – Data Access and Delivery Team
**Number of people**: (6)
**Location:** UC Davis Health

1. **Person submitting the application/ nomination**
	1. Cory Woods, Supervisor of Research Data Fulfillment, IT, UC Davis Health, staff.
	2. **Email address:** Woods@UCDavis.edu
	3. **The name of your organization:** UC Davis Health
2. **Award category:** IT Security
3. **Name of person, name of the team, or name of the project to receive the award:** Data Center of Excellence – Data Access and Delivery Team
4. **All project team members -**
5. Jodi Nygaard, Director of Data Access and Delivery, IT Data Center of Excellence, IT, UC Davis Health, Staff, jlnygaard@ucdavis.edu
6. Cory Woods, Supervisor of Research Data Fulfillment, IT, UC Davis Health, Staff, woods@ucdavis.edu
7. Ryan Davis, Compliance Program Executive Director, IT, UC Davis Health, Staff, rndavis@ucdavis.edu
8. Steve Covington, Programmer V, IT, UC Davis Health, Staff, scovington@ucdavis.edu
9. Brian Paciotti, Clinical Informatics Specialist V, IT, UC Davis Health, Staff, bmpaciotti@ucdavis.edu
10. Jeff Sterett, Supervisor of EMR Reporting, IT, UC Davis Health, Staff, jsterett@ucdavis.edu
11. **Which location was affected by the work?** UC Davis Health
12. **Summary** The Data Center of Excellence at UCDH created an SOP for the de-identification of ad hoc data sets. The resulting process provides transparency to data requestors, confidence to data provisioning teams acting as honest brokers, and enables us to continue to share learnings as we continue to improve the process and add automation and non-discreet data to the scope of the SOP.
13. **Narrative**

In May 2022, we began developing a standard operating procedure (SOP) that could be used by the Research Data Fulfillment team as well as other staff in the Data Center of Excellence to meet the needs of researchers as well as clinical operations. We began with documentation to outline in the scope of the SOP. In addition to documentation, several staff contributed from outside the DADT, including Institutional Review Board (IRB), Privacy and Compliance, Data Curation Team, IT Security, DADT supervisors, a policy analyst from the Clinical and Translational Science Center (CTSC) and staff that managed the deidentified Observational Medical Outcomes Partnership (OMOP) database. The first rough draft of the SOP was completed October 31, 2022.

The SOP starts by describing the stakeholders involved in the production of the SOP as well as the scope of the SOP. The next section discusses the Safe Harbor method and the 18 HIPAA identifiers that need to be removed or obfuscated from a de-identified data set. This portion also outlines the process that would be used for data to be de-identified and delivered to the data consumer. The difference between de-identified, anonymized, and limited data sets are explained in the following portion of the SOP.

The next section focuses on the technical process of de-identifying data sets. This part of the document provides examples of data that need to be excluding or obfuscated. Zip codes are the first example of data that needs to be obfuscated. To facilitate the process, csv files from the US Census Bureau were loaded into a database that is used to provide the data that will be included in the final data set. Dates and ages are discussed next. Several methods (showing year only, shifting the data randomly and providing intervals) are provided.

The process for building mapping tables is the next part of the SOP. Mapping tables contain the identified data as well as the de-identified data. These tables act as a key for linking PHI to de-identified data and can be used during the data validation process. This section also contains code (Oracle and MS SQL) for obfuscating an ID.

The next section is the most technologically dense part of the SOP. Code is provided that is used to build persistent and nonpersistent values for identified data such as patient ID, procedure ID, medical record number, etc. Following the ID obfuscation methods are code for applying random intervals to dates. The keys and intervals are maintained in the mapping tables mentioned above.

The next two chapters discuss the difference between a de-identified data set, an anonymized data set and a limited data set. The Quality Assurance process is articulated afterwards. Methods for resolving small cell counts, tracking and communicating the changes made to the data as well as a checklist are included in this section. The following part of the document details how to maintain the tables containing the Census Data. Lastly, roles and responsibilities and a quick reference table showing which data should be removed for each type of de-identified data set are the last part of the document.

The SOP is considered a living document and is revised based on what is learned from incoming de-identification projects. The document has been updated four times since the first draft, each time the SOP was used to fulfill a request for de-identified data. The first two projects utilizing the SOP pushed the boundaries of the SOP and were nearly out of scope as they included non-discrete data elements such as images or data that was considered sensitive and required further review. Insight from Privacy and Compliance staff was imperative to ensure the data sets had been fully de-identified but the data was still useful to the data consumer. Lessons learned from those projects were added to the document to facilitate future projects that may push the boundaries of the SOP. Establishing a learning and collaborative relationship with Privacy and Compliance staff assisted in revision to the document to ensure it is update and covers a variety of processes.

The SOP was made available to all staff in the Data Center of Excellence. The document provides enough detail so any staff can de-identify a data set. Resources are referenced in the document as well. In addition to making the document available to all staff, a tip sheet was created as a quick reference guide for data consumers to fully understand what will be provisioned to them and what will be removed or obfuscated. Training on de-identification and how to use the SOP will be conducted in the near future.

Developing the SOP led to the Research Data Fulfillment team being able to provide de-identified data sets, which is a success. Four data sets have been provided since the first draft of the SOP. Several additional data sets are currently under development and will be provide in the near future. The deidentified data sets have been used for research regarding cancer treatment, the impact of COVID on lung imaging, venous thromboembolism and a data set that will be used for machine learning. Without the SOP, the Research Data Fulfillment team would not have been able to provide de-identified data to the researchers without a full review by our Privacy Office each time. This SOP enables all provisioning teams to consult with Privacy when the bounds of the SOP are met, and what is learned is added to the SOP for future use.