
The main goal of data de-identification is to protect study participants' identity and privacy. There are several approaches used to de-identify data such as redacting, masking, and recoding direct identifiers to anonymize participants from a research study.

- Redacting may include removing or deleting an entire variable from a dataset (e.g., participant phone number), or a participant's data in a study (e.g., due to lack of consent for data sharing)
- Masking may include replacing data with an anonymized indicator or symbol (e.g., "James Smith admitted to hospital 2 hours after treatment" → "[Name] / XXXXX admitted to hospital 2 hours after treatment")
- Recoding may include applying a random code to anonymize data (e.g., changing site names from "Montgomery Hospital" to "1", and "York Hospital" to "2")

The general process when de-identifying data is to determine the data/variables to de-identify for personally identifiable information (PII) and protected health information (PHI), perform de-identification by applying various de-identification techniques, and document the de-identification procedures for traceability, replicability, and accountability.

The (C)DCCs are responsible for de-identifying data before submitting to the RADx Data Hub and should preserve as much data as possible to retain the scientific value, safeguard replicability, promote meaningful secondary use, and accommodate a wide range of potential future research by the scientific research community. The (C)DCCs de-identify RADx data of direct identifiers (e.g., names, addresses) and also based on the following guidelines:

- **Zip codes** – Consistent with Section 164.514(b) of the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule/Safe Harbor Method¹, the first 3-digits of a zip code are permissible to retain in the data. However, there are 17 restricted zip codes, which according to Census, have a population of 20,000 or fewer persons that should be changed to "000" in the data. The 17 restricted ZIP codes are: 036, 692, 878, 059, 790, 879, 063, 821, 884, 102, 823, 890, 203, 830, 893, 556, and 831.
 - Note: For studies that do not have any participants (e.g., wastewater studies), 5-digit zip codes are permissible.
- **Dates** – Full dates are allowed to be submitted to the RADx Data Hub with the date shifted by a constant number of days interval (e.g., add 5 days to all dates for a participant). If the population is very small (i.e., under 20 participants), all dates should be obscured to only contain the year.
- **Sites** – Sites within a study should be coded with a random ID to anonymize specific locations but still allow for site-level analyses.
- **Ages** – If a participant is less than one year old, the age will be listed as "0". For participants 21 to 89 years old, the age will be shifted to +/- 2 years. For participants 90 years and older, the age will be listed as "90".

¹ Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule (<https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html>)