Deidentification, Anonymization and Coded Data and Biospecimens in Human Tissue Research and the Impact on Biorepositories and Research and Discovery

Deidentification (either by anonymization or coding) is a method used to protect the privacy and confidentiality of individuals whose data or samples are procured, stored, or distributed by the Cooperative Human Tissue Network (CHTN), but they have distinct meanings and implications.

Deidentification involves the removal or alteration of personal identifiable information (PII) from data or samples to prevent the identification of individuals. This typically includes personally identifiable information such as names, addresses, social security numbers, and dates of birth (MM/DD/YYYY). Deidentified data or samples may still retain certain indirect identifiers such as age and gender, but they may be modified or generalized to reduce the risk of reidentification. The PII can be irreversibly anonymized (anonymization) or coded (reversibly anonymized).

Anonymization irreversibly removes all PII from the data and samples. This means that there is no way to link the data or samples back to the individual from whom they were obtained. Anonymization typically involves more extensive modification to the data or samples, such as removing all direct and indirect identifiers and ensuring that the remaining information cannot be used to identify the individuals by any means. This means that once the data or samples have been anonymized, they are no longer considered to be linked to any specific individual, and they can be used for research or other purposes without the need for consent. The anonymization process provides a higher level of privacy protection, but it may limit the utility of the data or samples for certain types of research that may generate results that can potentially be re-identified in certain circumstances (e.g., whole genome sequencing).

Coded samples are those in which PII (names, addresses, or medical record numbers) are replaced with a code that allows for tracking and management of the sample without revealing the identity of the individual. A key or code list is maintained separately from the samples, linking the codes to the individuals' identities. This allows researchers or authorized personnel to access additional information associated with the samples if necessary. Coded samples still retain the link to the individual's identity through the code, but the actual identifying information is kept separate and protected by an "honest broker". A code is sometimes also referred to as a "key," "link," or "map".

The "honest broker" is a neutral intermediary (person or system) between the patient's data and/or biospecimens being collected and the investigator. The CHTN acts as the honest

broker as it collects and collates pertinent information regarding the tissue source, replaces identifiers with a code, and releases only coded information to the investigator.

To learn more the following references discuss the concepts noted above in the context of human tissue biorepositories:

1. El Eman, K., Jonker, E., Arbuckle L., Malin, B. (2011). A Systematic Review of Re-Identification Attacks on Health Data. PLoS ONE, 6(12), e28071.
2. Ohm, P. (2009). Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization. UCLA Law Review, 57(6), 1701-1777. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1450006
3. Gkoulalas-Divanis, A., Loukides, G., Sun, J., Zhang, Y. (2016). Advances in Anonymization: Integrating Evolutionary Game Theory with Data Heterogeneity. IEEE Transactions on Knowledge and Data Engineering, 28(9), 2366-2379.
4. National Cancer Institute. (2018). Best Practices for Biospecimen Resources.
5. NIST Special Publication NIST SP 800-188 De-Identifying Government Datasets: Techniques and Governance (https://doi.org/10.6028/NIST.SP.800-188)