Duke | OFFICE *of* AUDIT, RISK & COMPLIANCE

IDENTIFIED DATA

For data to be used effectively and shared between entities, absent a consent or "explicit consent" (GDPR), all identifiable characteristics must be removed. Below are definitions covering degrees of identified data. Note, these definitions are often not very precise in that they usually include a catchall clause that will ultimately be resolved through litigation.

Degrees of Identified Data

Personal Data (GDPR): Any information that can be used to directly or indirectly identify a Data Subject (e.g., name, addresses, phone numbers, date of birth), or any information that relates to an identifiable person of someone who can be identified specific to physical, mental, genetic, physiological, economic, cultural or social identity.

Personally Identifiable Information (PII): This term has many different legal definitions. For a complete discussion of them, see https://en.wikipedia.org/wiki/Personally_identifiable_information.

Perhaps the most common is from NIST Special Publication 800-122, which defines PII as "any information about an individual maintained by an agency, including (1) any information that can be used to distinguish or trace an individual's identity, such as name, social security number, date and place of birth, mother's maiden name, or biometric records; and (2) any other information that is linked or linkable to an individual, such as medical, educational, financial, and employment information."

Individually Identifiable Health Information (IIHI, from HIPAA): This term does not include a definition of identifiable. Rather, it links to de-identification to indirectly obtain a definition.

Protected Health Information (PHI, from HIPAA): PHI is IIHI that is held by a HIPAA Covered Entity (e.g., healthcare provider, insurance company, etc.)

'pseudonymisation' (GDPR): "means the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person."

Anonymization (GDPR): is not defined in the GDPR, but can be inferred to be personal data that cannot be attributed to a specific data subject.

Identifiable private information (Common Rule): is used as part of the definition of a human subject (46.102). "Identifiable" means "the identity of the subject is or may readily be ascertained by the investigator or associated with the information". "Private information" is defined by example: "Private information includes information about behavior that occurs in a context in which an individual can reasonably expect that no observation or recording is taking place, and information which has been provided for specific purposes by an individual and which the individual can reasonably expect will not be made public (for example, a medical record)." In research, 'readily' has come to be interpreted as direct identifiers.

Coded Data (Common Rule): has been defined by OHRP as "identifying information (such as name or social security number) that would enable the investigator to readily ascertain the identity of the individual to whom the private information or specimens pertain has been replaced with a number, letter, symbol, and/or combination thereof (i.e., the code); and a key to decipher the code exists, enabling linkage of the identifying information to the private information or specimens." (https://www.hhs.gov/ohrp/regulations-and-policy/guidance/research-involving-coded-private-information/index.html)

k-anonymity: is a property of data by which the information for each person contained in the data release cannot be distinguished from at least k - 1 individuals whose information also appear in the release. (https://en.wikipedia.org/wiki/K-anonymity) K-anonymity is the idea from which the "rule of 10" for table cell sizes is derived. It is not as useful for anonymizing person level data.

De-identified (HIPAA): is accomplished by removing 18 identifiers or by having an expert certify that there is a "low probability" that an individual can be identified in the data. De-identified data can have a re-linking field as long as that field is not derived from identifiers in the data (e.g., "hash coded" re-linking fields are not allowed). HIPAA de-identified with a relinking field would be considered pseudonymous under the GDPR.

Limited Data Set (HIPAA): is a data set from which direct identifiers are removed. A "Data Use Agreement" is used control re-identification. Its use is limited to HIPAA operations, research, or public health. A relinking field may be included and it may be derived from direct identifiers. In research, a limited data set is very similar to Coded Data under the Common Rule.