SOP: Research Data Definitions				
NUMBER	DATE	AUTHOR	APPROVED BY	PAGE
PRDP-001	10/25/2020	Mary. M. A. Potter	Lisa M. Lee	1 of 7

1 PURPOSE

1.1 This standard operating procedure establishes common research data definitions.

2 REVISIONS FROM PREVIOUS VERSION

2.1 None

3 COMMON RESEARCH DATA DEFINTIONS

3.1 **Anonymous data:**

3.1.1 Anonymous data refers to data that were collected without any identifiers that could link the data back to an individual.

3.1.2 Examples of data in this category

3.1.2.1 Taste preference data collected at a grocery store

3.2 **De-identified data:**

3.2.1 De-identified data refers to data where all personally identifiable information (PII) have been permanently removed, severing identifiers from the observations. Any information that would allow someone to determine an individual's identity is removed from the data and not available to the research in the present or future. The primary reason for de-identifying data is to protect the identity of the individuals associated with the data to maintain privacy and confidentiality.

3.2.2 It is important to note that some datasets might indirectly reveal the identity of individuals even when the data seemingly contain no PII. For example, in areas with small populations information such as age, race, or occupation could allow a participant to be identified.

3.2.3 The most common strategies for de-identifying data are deleting all PII in a data file and either suppressing or masking certain variables so that the remaining information cannot be used to indirectly identify individuals.

3.2.4 Requirements to ensure de-identification vary. See References below for more information of the various regulations that restrict the use of PII and impose requirements for appropriately protecting such information.

3.2.5 Special note: Some de-identified datasets contain link identifiers or link codes, which are random or sequential numbers assigned to individual records that have otherwise had all PII removed. Link identifiers allow researchers to match data in two de-identified datasets without identifying participants.

3.2.6 Examples of data in this category

3.2.6.1 Public use datasets (e.g., national health surveys, census data, national economic and labor data)

3.2.6.2 Study data shared by other research teams

3.3 **Coded data:**

3.3.1 Coded data refers to a data that are initially collected with individual identifiers, which are subsequently replaced with a code (letter, number, or combination) that is unrelated to individually identifiable characteristics. A key to the code is maintained to allow the identifiable information to be linked back to the coded information. Coded data is generally considered "anonymous" until it is linked back up with the key.

3.3.2 Examples of data in this category

3.3.2.1 Data collected about students with ADHD that are identified with an alphanumeric code such as AB123, AB124, AB125, etc., where personal identifiers are linked with the code in a separate file.

3.3.2.2 Survey responses on drug use, where the responses are coded, and personal identifiers are linked with the code in a separate file.

3.4 **Personally Identifiable Information (PII)**

3.4.1 PII refers to any data or information about an individual that might reveal the identity or personal information or that could allow someone to indirectly identify a participant. Some identifiers might be indirect (e.g., zip code or date) that might need to be combined with other identifiers to be considered PII.

3.4.2 Obvious forms of PII include name, the names of parents or family members (including the maiden name of a student's mother), a household address, date of birth, social security number, identification numbers issued by government institutions or schools, and digital files such as photographs, videos, or audio recordings. Less obvious forms of PII might include biometric data (e.g., fingerprints or iris scans), geolocation data (e.g., real-time location data relayed by a smartphone), and metadata (i.e., "data about other data," such as location data about photographs, date of creation, and other data that are commonly embedded in digital files and photos).

PII is a legally defined concept used in federal and state regulations and reporting requirements. In the federal context, PII is defined in three statutes: Family Educational Rights and Privacy Act (FERPA); Children's Online Privacy Protection Act (COPPA); and the Protection of Pupil Rights Amendment of the Family Educational Rights and Privacy Act. In the state context, the Virginia Personal Information Privacy Act. [1]

3.4.3 Datasets that contain any of the following identifiers are considered to have PII. Some data elements (bolded below) are considered sensitive PII and require additional data protections.[2]

3.4.3.1 Name

3.4.3.2 Home address

3.4.3.3 Geographical subdivisions smaller than a state, including street address, city, county, precinct, zip code, and equivalent geocodes (note: the initial three digits of a zip code are not considered identifiable)

3.4.3.4 Full dates (month, day, and year) related to an individual, including birth date, date of death, and single year of age over 89 and all elements of dates (including year) indicative of such age (note: such ages and elements might be aggregated into a single category of age 90+)

3.4.3.5 Phone numbers

3.4.3.6 Fax numbers

3.4.3.7 Electronic mail addresses (e-mail)

3.4.3.8 Biometric identifiers, including finger and voice prints (audio recording)

---

[1] The Family Educational Rights and Privacy Act (FERPA) was passed in 1974 protects student information and defines the rights of students and their legally authorized representatives. The Children's Online Privacy Protection Act (COPPA) applies to information collected online through websites and apps from children under the age of 13, and the Protection of Pupil Rights Amendment of the Family Educational Rights and Privacy Act, which is intended to protect the privacy rights of students and parents. The Virginia Personal Information Privacy Act speaks directly to the requirements for the use and disclosure of PII. [1]

[2] See University Policy 1060, "Policy on Social Security Numbers" (https://policies.vt.edu/assets/1060.pdf) and the "Standard for High Risk Data Protection" (https://it.vt.edu/content/dam/it_vt_edu/policies/Standard-for-High-Risk-Digital-Data-Protection.pdf) for more information on the use of data elements in this category.

3.4.3.9 Full face photographic images and any comparable images (including video recording)
3.4.3.10 Social security number (SSN)
3.4.3.11 Passport number
3.4.3.12 Driver's license number
3.4.3.13 Student record number/identification number
3.4.3.14 Username for online or computer accounts
3.4.3.15 Bank account numbers
3.4.3.16 Credit and debit card number

3.4.4 PII risk level

3.4.4.1 The risk of the identifiers noted above is moderated by the specific data being collected as noted below.

3.4.4.1.1 **Benign** information about individually identifiable persons

3.4.4.1.1.1 Contains PII on human participants who have been given an assurance of confidentiality
3.4.4.1.1.2 Accidental disclosure is unlikely to result in harm to participants
3.4.4.1.1.3 The risks to the participant might be considered no greater than those associated with everyday life

3.4.4.1.2 **Moderately sensitive** information about individually identifiable persons

3.4.4.1.2.1 Contains PII on human participants who have been given an assurance of confidentiality
3.4.4.1.2.2 Could reasonably be expected to present a non-minimal risk of civil liability, moderate psychological harm, financial harm, or material social harm to individuals or groups
3.4.4.1.2.3 The risks to the research participant might be considered greater than those associated with everyday life

3.4.4.1.3 **Very sensitive** information about individually identifiable persons

3.4.4.1.3.1 Contains PII on human participants who have been given an assurance of confidentiality
3.4.4.1.3.2 Could cause significant harm to an individual if exposed, including, but not limited to, serious risk of criminal liability, serious psychological harm, financial harm, or other significant injury, loss of insurability or employability, or significant social harm to an individual or group
3.4.4.1.3.3 The risks to the research participant might be considered greater than those associated with everyday life

3.4.5 Examples of Data in this Category

3.4.5.1 Participant buying habits
3.4.5.2 Voter registration analysis including demographic information

3.4.5.3 Analysis of student performance on exams

3.5 **Sensitive Personally Identifiable Information Data (SPII)**

3.5.1 SPII refers to information that if lost, compromised, or disclosed could result in substantial harm, embarrassment, inconvenience, or unfairness to an individual. It includes any information that could be used by bad actors to conduct identity theft, blackmail, stalking, or other crimes against an individual.

3.5.2 SPII requires stricter handling guidelines because of the increased risk to an individual if the data are inappropriately accessed or compromised. SPII is subject to same federal and state regulations and reporting requirements as PII, including the Protection of Pupil Rights Amendment of the Family Educational Rights and Privacy Act (FERPA) and the Virginia Personal Information Privacy Act.

3.5.3 SPII can consist of stand-alone data elements, including:

- 3.5.3.1 Social security number
- 3.5.3.2 Biometric information
- 3.5.3.3 Bank account number
- 3.5.3.4 Passport information
- 3.5.3.5 Health care related information
- 3.5.3.6 Medical insurance information
- 3.5.3.7 Student information[3]
- 3.5.3.8 Credit and debit card number
- 3.5.3.9 Driver's license and state ID information
- 3.5.3.10 These additional data elements can make PII more sensitive: citizenship or immigration status, ethnic, religious, sexual orientation, or activities that can have a negative effect on reputation in conjunction with the identity of an individual (directly or indirectly inferred), are considered SPII.

3.5.4 When determining the sensitivity of PII, researchers should evaluate the sensitivity of each individual PII data element, as well as the sensitivity of the data fields together (e.g., an individual's SSN, medical history, or financial account information is generally considered more sensitive than an individual's phone number or zip code).

3.5.5 The sensitivity of PII might be greater when combined with other information (e.g., name and credit card number are more sensitive when combined than apart) or in certain contexts, such as on a clinic's patient list.

3.5.6 Examples of data in this category

- 3.5.6.1 Data about individuals with a gambling addiction
- 3.5.6.2 Studies about incarcerated individuals
- 3.5.6.3 Data on immigrant populations

3.6 **Private information**

3.6.1 Private Information is information associated with individuals or groups of individuals that could reveal details of their lives or other characteristics that could cause physical, financial, social, or other harm to them.

3.6.2 Private information is typically a classification of information that individuals use for themselves. It is a broad and general term that is more ambiguously used than other privacy terms. For example, the combination to a bank safety deposit lock is private, but the combination number itself does not point to any specific individual. As another

[3]Student health records at postsecondary institutions receiving funding from the U.S. Department of Education (DoED) are considered "education records" under the US Family Educational Rights and Privacy Act (FERPA).

example, some individuals consider how they voted in presidential elections to be private information that they do not want any others know.

3.6.3 Individuals often consider PII to be a type of private information, and personal information could also be private information. For utilities, market data that includes information about a negotiated price for a customer is likely considered by the customer to be private information; they might not want their friends, neighbors or the general public to see this information.

3.7 **Protected Health Information (PHI)**

3.7.1 PHI refers to health data created, received, stored, or transmitted by health care entities (and their business associates) in relation to the provision of health care, health care operations, and payment for health care services. PHI does not include individually identifiable health information of persons who have been deceased for more than 50 years.

3.7.2 PHI applies to research that uses, creates, or discloses PHI that enters the medical record or is used for health care services, such as treatment, payment, or operations. PHI is used in studies that review existing medical records to collect research data, such as retrospective chart review; or PHI entered into the medical record that was produced in the course of a research study, such as diagnosis of a health condition or evaluation a new drug or device. For example, sponsored clinical trials that submit data to the U.S. Food and Drug Administration involve PHI and are therefore subject to HIPAA regulations.

3.7.3 PHI includes individually identifiable health information, including demographic data, medical history, test results, insurance information, and information used to identify a patient or provide services or health care coverage. There are 18 named identifiers that can be used to identify, contact, or locate a person. If health information is used with any of these identifiers it is considered identifiable and might be subject to regulatory protections. If all of the 18 identifiers are removed the data is no longer considered to be PHI.

3.7.4 The following is a list of the specific 18 identifiers that when obtained from a health care entity are considered identifiable PHI:

- 3.7.4.1 Names (full or last name and initial)
- 3.7.4.2 All geographical identifiers smaller than a state, except for the initial three digits of a zip code if, according to the current publicly available data from the U.S. Bureau of the Census: the geographic unit formed by combining all zip codes with the same three initial digits contains more than 20,000 people; and the initial three digits of a zip code for all such geographic units containing 20,000 or fewer people is changed to 000
- 3.7.4.3 Dates (other than year) directly related to an individual
- 3.7.4.4 Phone numbers
- 3.7.4.5 Fax numbers
- 3.7.4.6 Email addresses
- 3.7.4.7 Social security number
- 3.7.4.8 Medical record number
- 3.7.4.9 Health insurance beneficiary numbers
- 3.7.4.10 Account numbers
- 3.7.4.11 Certificate/license numbers
- 3.7.4.12 Vehicle identifiers (including serial numbers and license plate numbers)
- 3.7.4.13 Device identifiers and serial numbers
- 3.7.4.14 Web Uniform Resource Locators (URLs)
- 3.7.4.15 Internet Protocol (IP) address numbers
- 3.7.4.16 Biometric identifiers including finger, retinal, and voice prints

3.7.4.17 Full face photographic images and any comparable images

3.7.4.18 Any other unique identifying number, characteristic, or code except the unique code assigned by the investigator to code the data

3.7.5 Examples of data in this category

3.7.5.1 Data obtained from a health care system on individuals being treated for diabetes

3.7.5.2 Data obtained from tumor/cancer registries

3.7.5.3 Data obtained from an autism treatment center about children with Asperger's syndrome

3.7.6 PHI is a legally defined concept used in federal and state regulations and reporting requirements. In the federal context, PHI is defined in two statutes: The Health Insurance Portability and Accountability Act of 1996 (HIPAA) and the Health Information Technology for Economic Clinical Health Act (HITECH Act). In the state context, the Virginia Personal Information Privacy Act speaks directly to the requirements for the use and disclosure of PHI.

3.7.7 Research studies that use health-related information that are not associated with or derived from a health care service event (treatment, payment, operations, medical records) are not considered PHI. HIPAA does not apply to "research health information" (RHI) that is kept only in the researcher's records; however, other human subjects protection regulations still apply. Some genetic and genomic research can fall into this category, such as the search for genetic markers for a particular condition. In contrast, genetic testing, when entered into a medical record as part of diagnosis for the treatment of a patient, is considered PHI and is subject to HIPAA regulations.

3.7.8 PHI applies to health records but not to student health information held by an educational institution or health information held by a health care entity related to its role as an employer.

3.8 **Other Sensitive Research**

3.8.1 Researchers might also obtain, store, use, or share sensitive information that does not relate to human subjects. Examples include:

- Potential intellectual property
- Proprietary information subject to confidentiality requirements
- Information with export control/national security restrictions

Questions related to intellectual property should be directed to LINK, LICENSE LAUNCH as well as review of Policy on Intellectual Property (Policy 13000). Questions related to export control, including international travel, or national security restrictions should be directed to the Office of Export and Secure Research Compliance, as well as review of Export Control, Sanctions, and Research Security Compliance Policy (Policy 13045).

4 RESPONSIBILITIES

4.1 PRDP will underline the terms defined in this standard in related standard operating procedures.

4.2 Related Standard Operating Procedures

4.2.1 PRDP-002 Personally Identifiable Data Research Protections

5 PROCEDURE

5.1 None

6 MATERIALS

6.1 None

7 REFERENCES

7.1 Children's Online Privacy Protection Act (COPPA) protects the privacy of children under the age of 13 by requesting parental consent for the collection or use of any personal information of the users.

7.2 Fair Credit Reporting Act (FCRA) U.S. Federal Government legislation enacted to promote the accuracy, fairness, and privacy of consumer information contained in the files of consumer reporting agencies.

7.3 Family Educational Rights and Privacy Act (FERPA) US Federal law that governs the access to educational information and records by public entities such as potential employers, publicly funded educational institutions, and foreign governments.

7.4 General Data Protection Regulation (GDPR) provides the rules relating to the protection of how personal data is processed and moved. This Regulation protects fundamental rights and freedoms of individuals, specifically their right to the protection of personal data.

7.5 Gramm-Leach-Bliley Act (GBLA) requires companies that offer consumers financial products or services like loans, financial or investment advice, or insurance to explain their information-sharing practices to their customers and to safeguard sensitive data.

7.6 Health Insurance Portability and Accountability Act (HIPAA) required the Secretary of the U.S. Department of Health and Human Services (HHS) to develop regulations protecting the privacy and security of certain health information.

7.7 Virginia Privacy Laws provide required safeguards to protect the use of consumer and patient data.