Recommendations: Securing Human Subject Research Data

Data breach represents a significant risk to participants in human subject research which must be minimized. The potential magnitude of risk involved depends on the sensitivity of the data. The probability of data breach due to increased occurrence can be minimized through implementation of data security measures. In general, data that are of low or intermediate sensitivity can be secured by researchers with minimal training. For research that will involve highly sensitive or legally protected data, a data security plan should be developed in collaboration with your college or department IT personnel.

The table below provides a general overview of security level and security control recommendations for working with data collected from human subjects research. A more detailed description is provided for each security level on subsequent pages. For information on required security controls, please see IT Services Guidance document "Research Data Security Protections" and Review WSU's EP8 policy.

Level	Security of Data	Brief Examples	Security Controls
1	Low-Public	Publicly available data, anonymized data, or de-identified data that cannot be re-identified by the researcher.	Password protected documents and files. Locked filing cabinets and offices for hard copy data.
2	Intermediate-Protected	Identifiable data (e.g. surveys, field notes video recordings) which, if disclosed would NOT put subjects at risk (reputational, employability, legal, embarrassment).	Encrypted files and flash drives. Consult IT as needed.
3	High-Restricted	Sensitive data such as medical records with identifiers. Legally protected data (e.g. HIPAA /FERPA) when appropriately protected. Research involving criminal activity that is protected by Certificate of Confidentiality (CoC).	Same as Level 2 but verified by college IT personnel. Encrypted laptops or work stations. Certificate of Confidentiality (CoC) when appropriate
4	Very High Restricted	Research records with identifiers of self-disclosed criminal activity or mental health issues. Research that requires or documents violation of federal or state law (e.g. consumption of cannabis). Data that contractually or legally protected or deemed classified.	Same as level 3 with additional precautions (e.g. non-networked work-station) as determined by college IT Director or designee.

RESEARCH DATA SECURITY LEVELS: Definitions and Examples

Level 1 Low Security – Public or Anonymous Data

IT Definition of Public: Information that is currently released or approved to be released to the public without restriction by the appropriate information owner. Information in this classification does not need protection from unauthorized access or disclosure; however, there may be requirements to protect the integrity and availability of data in this classification. Examples of public information are employee directory information, public University outreach and research publications, press releases, and information on the public WSU website.

Research Definition of Public: Information that is anonymous and not sensitive; or received as de-identified and will remain so; or data that is available from a public information source. These data have little or no sensitivity and disclosure of it would have little or no risk of physical, psychological, social, economic, legal, or educational advancement harm to individuals.

Example	Representative Case
Information that was identified, but the researcher only has access to de-identified information.	The researcher is collaborating with the Washington Department of Education (WDE). WDE will send paper surveys out to math instructors at all Oregon community colleges. The survey will request information about what changes they would like their administrators to make with regard to curriculum. WDE will remove any identifying information before providing it to the WSU researcher, and will sign an agreement that it will never provide the identifiers to WSU or the researcher.
Publicly available data.	Telephone directory information, zip code maps.
Information from public web pages that do not require payment or a password to access.	www.wsu.edu www.census.gov
Data that is collected anonymously.	An online Qualtrics survey is sent to directors of drug rehabilitation programs. The survey asks about their opinions regarding counseling practices. Data anonymization is turned on in Qualtrics so that no IP addresses are captured.
Data that is collected and linked with no identifying coding at any time, such as where persons are given a random code used to link data collected at different time points, rendering the collected data not linked to an identifier at any time.	A researcher wants to do a pre-survey, an intervention, and then a post-survey. She uses a code to link the two surveys that does not allow the subject to be identified. For each stage of collection, she uses the subject's favorite number, their favorite color, and the first number of their street address as a code to link the information together.

De-identified data that is not layered with other data to make it identifiable, in a very large aggregate (generally 30,000 persons or more).	40 year old males in Washington State. Even if this information was layered with 20 year old males in Washington State, there would be no overlap to identify the individuals.
Aggregated de-identified Personal Health Information.	Of those admitted to Pullman Hospital, 40% of males have a cardiovascular risk factor.
Public observation of a non-vulnerable population.	A researcher counts how many individuals wear Cougar shirts on Non-Game Days vs. Game Day.

RESEARCH DATA SECURITY LEVELS

Level 2 Intermediate Risk – Internal Data

IT Definition of Internal: Information that is intended for official WSU business purposes only. This information may be made available to authorized University personnel with a legitimate need in support of the performance of their assigned roles/duties, and may be released to authorized University affiliates or third parties with approval from the appropriate information owner, or as required by law. It is not appropriate for information in this classification to be made available to the general public. Unauthorized access, disclosure, or loss of integrity or availability of this classification of information could result in some harm to the University or to individuals. Examples of internal information may include information concerning various University business transactions, operations, and strategies and methods that may be considered to provide a competitive advantage.

Research Definition of Internal: Identified benign or innocuous data; or identified private or confidential or sensitive data that if released would not have significant impact on the individual or University; or anonymous but sensitive data (due to risk of re-identification – for example, by layering, signed consents, payments, or small population size). Research data that was originally identified, but will be de-identified. Research data that you would share with your team or a colleague, but not outside that select group. Data that is not required by contract or regulations to be secured in a particular fashion.

Example	Representative Case
De-identified student grades or individual work.	A professor copies student essays for his research, but then removes all names and other identifiers such as student ID numbers prior to analysis.
Data that is collected and linked with confidential coding where the master list is kept separate from the data.	A researcher has a list of subjects with a number next to each name. The dataset contains only the number and the rest of the data. The list and dataset are kept in separate files.
De-identified innocuous or sensitive data that is layered with other data to make it likely to be identifiable.	Data Set 1: 40 year old males. Data Set 2: Specific street blocks in Pullman, WA. Data Set 3: Unmarried, Data Set 4: Owns a cat. When combined, the dataset could likely pinpoint with a fair amount of accuracy the individuals who meet this criteria.
De-identified Personal Health Information.	The completely de-identified patient charts of all 40 year old males with a specific cardiovascular risk factor who take statins in one hospital.
Identified data where disclosure may not lead to criminal or civil liability, or be damaging to financial standing, employability, or reputation.	An identified response to questions on political opinions. While the data is not sensitive, it is primarily considered confidential and/or private.
De-identified, but sensitive information.	De-identified criminal records.

	De-identified psychiatric commitment records.
Unpublished research work and intellectual property.	A researcher has a draft version of a paper of her research results.
Anonymous but sensitive data.	An online Qualtrics survey will be sent out to college instructors across the country. The survey will include questions regarding the barriers that they face in working with their administration. Data anonymization is turned on in Qualtrics so that no IP addresses are captured. A drug rehabilitation clinic posts a flier in their facilities requesting that clients contact the researcher to participate in an interview where only hand-written notes will be taken. The interview will ask questions about how many times the client has started a rehabilitation program, how many times they have completed programs, and their opinion about different counseling styles that they have experienced.
An innocuous interview that is recorded, but then the recording is deleted after transcription.	A researcher interviews farmers about their favorite varieties of lentils to grow. He transcribes the interview and then deletes the original recording. He keeps the identifier along with the transcription.
Research comparing educational practices in a commonly accepted educational setting.	A researcher compares two different approaches to teaching a mathematical concept to 3rd grade students.
Public behavior observation of vulnerable populations.	A researcher counts how much time children play on swings versus slides in a public park. Because children are a vulnerable population, additional care is taken with the data.
A benign behavioral intervention in which identity is recorded.	A researcher has identifiable subjects take two similar tests with and without classical music playing in the background to compare results. While the data is not harmful, it is research data and should be appropriately protected from theft and misuse.
Publicly available private information or biospecimens.	A researcher uses data from a publicly available data repository that requires no passwords or other criteria (such as credentials) to obtain.
Secondary research information recorded by a researcher without identifiers.	A researcher has access to medical charts, but does not record subject identifiers and will not contact the subjects.

Taste and food quality evaluation.	A researcher has subjects try two different apple samples to determine which is preferable. Identities of subjects are recorded. While the data is not harmful, it is research data and should be appropriately protected from theft and misuse
Identified data where disclosure is innocuous.	Instructors of community college math programs are asked to participate in interviews about their opinions regarding different educational practices.

RESEARCH DATA SECURITY LEVELS

Level 3 High Risk – Confidential/Restricted Data

IT Definition of Confidential: Information that is specifically protected by law, contracts, third-party agreements, or for other University business reasons as established by the appropriate information owner. Access may be granted to this classification of information by the appropriate information owner to only authorized personnel with a legitimate need-to-know. Confidential information may be released to authorized University affiliates or third parties only with explicit approval from the appropriate information owner, or as required by law. Unauthorized access, disclosure, or loss of integrity or availability of this information could cause significant harm to the University and its operations, assets, or individuals. Information in this category may include employee personnel records, financial information, donor information, intellectual property, attorney/client privileged information, information regarding critical infrastructure of physical structures and assets, and the security and infrastructure of information technology systems.

Research Definition of Confidential: Private or sensitive data which if released could have genuine impact and result in serious, long-lasting, or irreversible harm to the University or individuals. Individually identifiable information that, if disclosed, could reasonably be expected to be damaging to a person's reputation or to cause embarrassment. The disclosure of this data could place the subjects at risk of criminal or civil liability or be damaging to the subjects' financial standing, or employability. This includes identifiable data that is confidentially coded and if linkage is kept, or data that is collected with identifiers. This data should be protected by a Certificate of Confidentiality (CoC) as appropriate.

Example	**Representative Case**
Taste and food quality evaluation of legal, but regulated, substances.	A researcher has subjects try two different wine samples to determine which is preferable. Identities of subjects are recorded and a copy of the subject's drivers' license is obtained as proof of legal age.
Research involving identifiable educational tests (cognitive, diagnostic, aptitude, achievement) survey procedures, interviews, or observation of public behavior with visual or auditory recording, and the collected information could reasonably place the subjects at risk of criminal or civil liability or be damaging to the subjects' financial standing, employability, educational advancement, or reputation. When protected by an NIH Certificate of Confidentiality (C0C).	A researcher conducts an identifiable survey of subjects' sexual habits. A researcher interviews an identifiable subject about their illegal drug use. A researcher videotapes persons committing a criminal offense. A researcher gives a subject psychological tests to determine if the subject is autistic.

Identified student grades or individual work, admission application, class schedule, or information on individuals who have requested a FERPA block.	A researcher uses student test results and class schedules to determine if student achievement is better in the morning or afternoon.
Identified data, including screening data, that is sensitive and where disclosure would lead to criminal or civil liability, psychological harm, loss of insurability, or be damaging to financial standing, employability, or reputation of an individual or group. When protected by an NIH certificate of confidentiality.	An identified response to questions about personal criminal activity or specific lies a subject tells on employment applications. Genetic health risk of a group. Identifiable genotyping. Self-reported mental health or genetic information, drug or alcohol abuse, or illegal behaviors. An online Qualtrics survey will be sent out to college instructors across the country. The survey will include questions regarding the barriers that they face in working with their administration. Data anonymization is turned off in Qualtrics so that IP addresses are captured. A drug rehabilitation clinic sends out a Qualtrics survey to previous clients on behalf of the researcher. The survey includes questions about past illegal drug use and private health information. Data anonymization is turned on in Qualtrics so that no IP addresses are captured. The researchers will not have access to names or email addresses of clients, however, demographic questions asked are date of birth, gender, ethnicity, and zip code.
De-identified sensitive data that is layered with other data to make it likely to be identifiable.	A researcher combines Data Set 1: 40 year old males. Data Set 2: Specific street blocks in Pullman, WA. Data Set 3: Criminal Record Data Set 4: IP Address
Identifiable criminal records, law enforcement information, or other sensitive data.	A researcher accesses identifiable psychiatric commitment records.
Information protected by a NIH Certificate of Confidentiality.	An NIH funded researcher is asking identified subjects about their cannabis usage, and has received an NIH Certificate of Confidentiality.

Identifiable signed consent forms for research that is on sensitive, damaging, or illegal aspects of life.	A researcher has signed consent forms from subjects agreeing to an interview about their experience in an alcohol rehabilitation program, or domestic violence treatment, or heroin usage.
Contracts, Memorandums of Understanding, Non-Disclosure or other Agreements or Non-Public Contracts.	A researcher has a Data Use Agreement allowing him access to unredacted Section 8 Housing program records.
Personal information protected under state, federal, or foreign privacy laws.	A researcher is collecting survey results from persons residing in a GDPR country. A researcher is interviewing subjects in Brazil, which has similar privacy standards to the GDPR although it is not a GDPR country. A researcher wants to conduct an online survey of 12 year olds who are protected by COPPA.
Clinical studies of drugs and medical devices.	A researcher wants to test the effects of grapefruit juice on medication absorption, and will be collecting a self-reported medical history along with multiple biological samples during the study.
Identifiable biospecimens.	A researcher collects blood samples and hair clippings.
Identifiable data collected through noninvasive procedures.	A researcher has a subject wear a fitness monitoring device or ECG cap.

RESEARCH DATA SECURITY LEVELS

Level 4 Very High Risk–Regulated or Contractually Protected Data

IT Definition of Regulated: Information that is specifically protected by federal, state, local, or industry policies and/or laws and regulations, for which strict protection, use, and handling requirements are dictated. Access may be granted to this classification of information by the appropriate information owner to only authorized personnel with a legitimate need-to-know. This information may be released to affiliates or groups outside of the University community only with explicit approval from the appropriate information owner, or as required by law. Unauthorized access, disclosure, or loss of integrity or availability of this information could cause serious harm to the University and its operations, assets, or individuals. Data in this classification may be exempt from public records or other legal requests.

Research Definition of Regulated/Contractually Protected: The release of this data would likely result in serious, long-lasting, or irreversible harm to the University or individuals. Information that could cause harm to an individual if disclosed, including, but not limited to, risk of criminal or civil liability, psychological harm or other injury, loss of insurability or employability, or social harm to an individual or group. Research data that is regulated by a federal agency or requires special data protection due to contractual arrangements. This data may have a statutory requirement for notification to affected parties if a breach of confidentiality occurs. FOIA exemptions and exceptions are likely to apply.

Example	Representative Case
Consumption of illegal substances.	A researcher has subjects ingest cannabis, which while legal in Washington State is illegal under federal law. The researcher takes blood samples before and after consumption and codes them to a master list which is kept separate from subject names.
HIPAA protected information.	A researcher wants to receive medical records from a covered entity. The records are protected by HIPAA and require specific IT protections.
Clinically validated results.	A researcher has testing done by a CLIA certified laboratory and provides a copy of the results to the subject's doctor.
A subject gives a researcher permission to access financial, medical, or other sensitive personal data.	The researcher requests to access the subject's medical record, employment history, and to measure radon levels in the subject's home.
Information covered by a regulation, agreement, or national security that requires that data be stored or processed in a high security environment.	A researcher is granted access to the State of Washington Department of Social and Health Services' files, but the contract specifies certain data security requirements.

	A researcher is working with the Department of Defense on a contract that requires certain security clearances.
Individually identifiable genetic information.	Whole genome sequencing. Individually identifiable sensitive genotype information.
Export Controlled Information.	A researcher is working with information that may not be released to a specified foreign country.
Individually identifiable sensitive data.	Student loan application information including financial aid and grant information. Legal presence status in the United States. Information that can lead to identity or financial theft. (Social security numbers along with name and birth date, bank account numbers, copy of a passport, etc.).