Cornell University
Office of
Research Integrity and Assurance

DISCLAIMER: This document has not been updated to reflect regulatory changes effective January 21, 2019.

Human Research Participant Protection Program

Guidance on Research Involving Only Existing or Secondary Data, Documents or Records

I. SUBJECT

This guidance is designed to assist investigators planning research that involves no interaction or intervention with individuals, and uses only existing or secondary data about living individuals.

By federal regulation, existing (or secondary) data are defined as data that existed 'before the research is proposed to an institutional official or the IRB'.[1] This provision is frequently interpreted as data that were 'on the shelf' at the time the protocol was written.[2] Existing data include both those provided to the investigator from any source; and those already in the possession of the investigator.[3] Investigators planning to use biological samples for research should also consult IRB Policy #19. [See Appendix A for other definitions.]

Data providers and funders of sponsored research may require IRB approval or certification of exemption from IRB review, which takes precedence over this guidance.

II. GUIDANCE

Researchers should use this document and the Decision Charts in Appendix B, to make an initial determination for whether their project using existing or secondary data requires IRB approval or certification of exemption from IRB review.

A. Secondary data analysis that does not require review by the IRB office

Use of the following types of public data does not constitute federally-regulated human subjects research and requires no further action with the IRB office.

1. Data that are **not about living individuals**
   Examples: Historical records of deceased individuals, death records, historical archives
2. **Publically available data**; i.e. data with identifiable but **not private** information.
   Examples: Individual public records such as address, phone number, property value; data found on unrestricted websites, in publications, phone books, political campaign contributions, or obtained through a Freedom of Information Act request.

   Note1: Even though the use of such data is excluded from IRB review, researchers must ensure that the data is obtained under conditions allowed by the data provider, and exercise best practices appropriate to the field.

   Note 2: A project that **merges public datasets with other datasets containing private information** may enable the identification of individuals and requires IRB review. Investigators using such merged data must submit an Initial Application for Approval.

B. Secondary data analysis that may require IRB review

1. **Non-public, de-identified or anonymous data**
   Use of the following types of non-publicly available datasets does not constitute federally-regulated human subjects research, if **the provider of the data is NOT involved in the design, conduct or reporting of the research**, including sharing any authorship rights:

   - Datasets that are anonymized (see definition in Appendix A) or
   - Datasets that include coded private information, such that the investigator has no access to identifiable information[1]

   In such cases, no further action with the IRB office is required.

   Investigators are advised to have on file a confidentiality statement from the data provider stating that identifiers were not included in the dataset that they received. [A sample confidentiality statement for de-identified data without the key is included in Appendix C.]

   Example: De-identified data given by a colleague or provided by a data provider where no collaboration is expected.

2. **Non-public, identifiable data, but where researcher does not record, retain or use identifiers**

   Use of datasets that contain private, identifiable data, but from which **no identifiable information will be recorded, retained or used by any member of the research team** in a manner that allows the direct or indirect identification of individuals, may be eligible for exemption from IRB review. Investigators are required to submit a Request for Exemption from IRB Review.

   Investigators should include a data security plan that clearly describes measures to ensure that individuals cannot be identified, including the process for de-identifying data, and plans for the disposition of the identifying information. More information on data security plans is on the IRB website at www.irb.cornell.edu.

3. **Non-public, identifiable data where researcher has access to and records private identifiable information**
   Use of non-publicly available data when members of the research team or their collaborators have access to and plan to use private identifiable information about living individuals, is considered human subjects research and requires review by the IRB. Investigators using such data are required to submit an Initial Approval Request .

---

[1] Cornell investigators may not have direct access to identifiers or to a code linking the specimens/data to identifiable living individuals. Some examples of how this access is prevented, are below:

- the key to decipher the code is destroyed before the research begins; or
- the investigators and the holder of the key to the code enter into an agreement preventing the release of the key to investigators under any circumstances; or
- there are IRB-approved written policies or legal requirements in place preventing/prohibiting the release of the key to Cornell investigators

Example: Birth records matched to other data using individual identifiers.

Due to the potentially sensitive nature of these data, a data security plan is required. More information on data security plans is on the IRB website at www.irb.cornell.edu.

C. Two Special Cases:

1. **Data available through the New York Census Research Data Center (NYCRDC) at Cornell:**

   Use of Census data at NYCRCD has been granted an Exemption from IRB review (#1106002306). This dataset has private, identifiable information but no identifiers are recorded or used by investigators. Investigators access this data under a data use agreement executed by Cornell Restricted Access Data Center (CRADC). Investigators using *only these data* do not need to take any other action with the IRB office.

2. **HIPAA-protected data:**
   Investigators on the Cornell Ithaca campus may use protected health information (PHI) covered by HIPAA regulations only under the terms and conditions outlined by the Covered Entity (e.g. a health care provider or institution) that is providing the data. Investigators using PHI are required to submit an Initial Approval Request to the IRB office, a letter of authorization from the covered entity providing the PHI, a data use agreement negotiated and executed by Cornell's Office of Sponsored Programs (OSP), and a data security plan that describes how the proposed data privacy protections meet HIPAA standards. Detailed guidance on the use of HIPAA protected data for research at Cornell is available on the IRB website.

III. WRITTEN AGREEMENTS FOR RESTRICTED ACCESS OR LICENSED DATA

An investigator planning to obtain data under contractual terms, such as a restricted access data agreement or a licensed data agreement, must contact the Office of Sponsored Programs (OSP). OSP will negotiate the terms of that agreement with the data provider on behalf of the University. Cornell investigators are not authorized to sign data use agreements themselves.

Investigators must include with their IRB application 1) a copy of the fully executed data use agreement and 2) an approved data security plan. When the sponsor or provider of the data requires IRB approval before the data agreement can be finalized, the IRB office can issue a conditional approval for the project.

IV. References

1. USDHHS, Office of Human Research Protections (OHRP). Human Subjects Regulations Decision Charts, September 24, 2004. http://www.hhs.gov/ohrp/policy/checklists/decisioncharts.html#c5
2. National Institutes of Health, Office of Extramural Research. Frequently asked questions from applicants, Feb. 2010. http://grants.nih.gov/grants/policy/hs/faqs_aps_definitions.htm#285
3. US-DHHS, Office of Human Research Protections (OHRP). Guidance on Research Involving Coded Private Information or Biological Specimens, Oct. 16, 2008. http://www.hhs.gov/ohrp/policy/cdebiol.html

Appendix A

Terms and Definitions (also see the Cornell IRB glossary)

- Research is a 'systematic investigation, including research development, testing and evaluation, designed to develop or contribute to generalizable knowledge.'[1]
- Human subjects or human participants are 'living individuals about whom an investigator (whether professional or student) conducting research obtains:
    1. data through intervention or interaction with the individual or
    2. identifiable private information.'[1]
- Human subjects research is 'research' involving 'human subjects' (as defined above).
- 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 (e.g., a medical record). Private information must be **individually identifiable** (i.e., the identity of the subject is or may readily be ascertained by the investigator or associated with the information)'.[1]
- Existing (or secondary) data is defined as data that existed 'before the research is proposed to an institutional official or the IRB'. [2] This provision is frequently interpreted as data that were 'on the shelf' at the time the protocol was written.[3] Existing data include both those provided to the investigator from any source, and those already in the possession of the investigator.[4]
- Publicly available (or public use) data files are generally 'prepared by investigators or data suppliers with the intent of making them available for public use'[5] , although a dataset can be made public after the fact by following careful de-identification schema. Guidance on de-identification methods is provided in Reference #5 below. The data available to the public are not individually identifiable or maintained in a readily identifiable form.
- De-identified (or coded) data come from a dataset that contains:
    - Identifying information that would enable the investigator to directly ascertain the identity of the individual to whom the private information pertains has been replaced with a number, letter, symbol, or combination thereof (i.e., the code); ***AND***
    - a key to decipher the code exists, that enables the direct linkage of the identifying information to the private information.' [4]
- Anonymized data are de-identified data for which a code or other link back to private information no longer exists. An investigator would not be able to link anonymized information back to a specific individual.
- Restricted access data must be used with appropriate confidentiality protections as specified in a formal written **data use agreement** between the University and the data provider.
- Licensed or limited use data are used within an agreement that protects the provider's intellectual property (e.g., copyright), and may or may not include adequate steps to assure the confidentiality of participants.

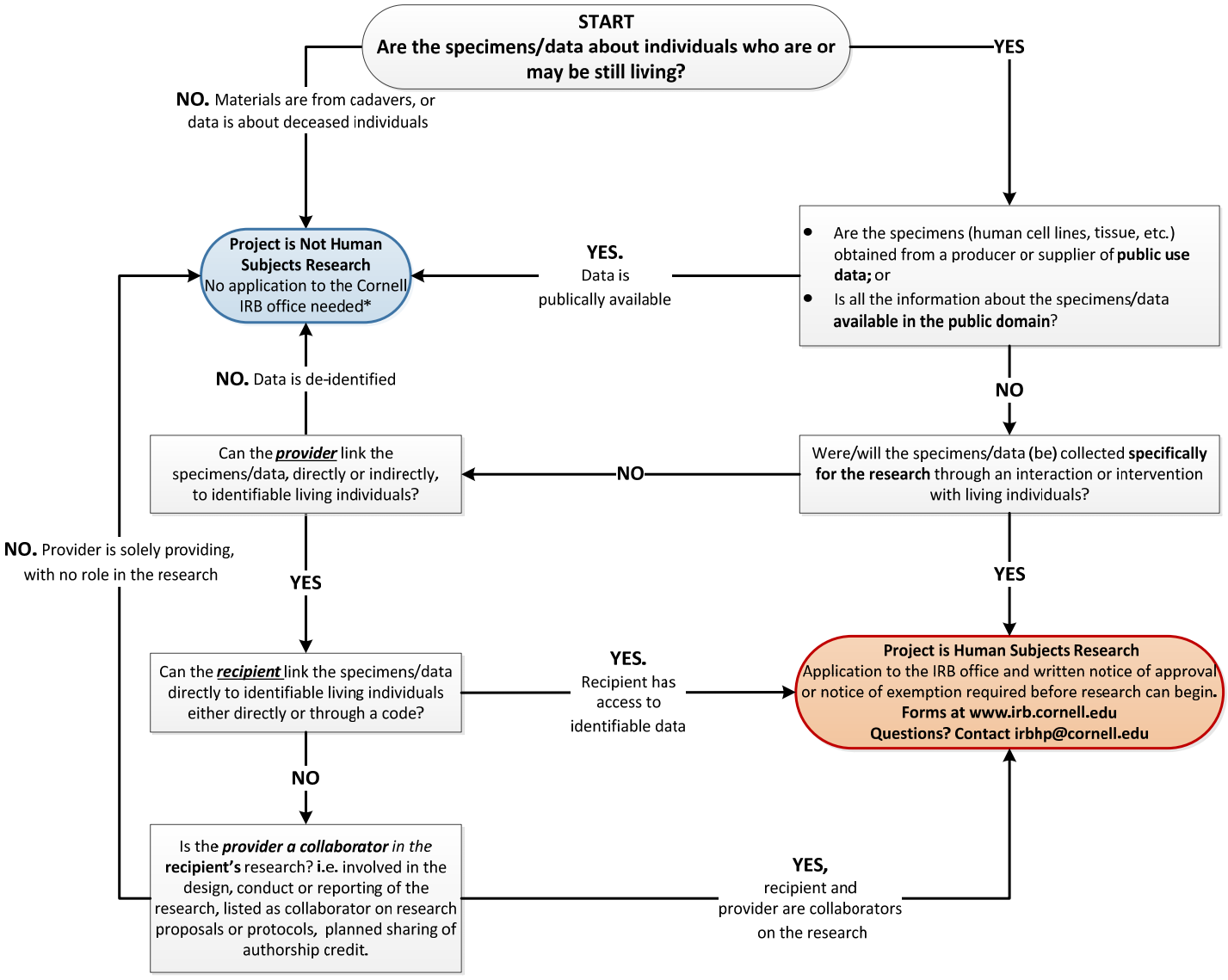
- Anonymous data were originally collected without identifiers or on unknown persons.

References

1. 45 CFR 46.102 http://www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.html
2. US-DHHS, Office of Human Research Protections (OHRP). Human Subjects Regulations Decision Charts, September 24, 2004. http://www.hhs.gov/ohrp/policy/checklists/decisioncharts.html#c5
3. National Institutes of Health, Office of Extramural Research. Frequently asked questions from applicants, Feb. 2010. http://grants.nih.gov/grants/policy/hs/faqs_aps_definitions.htm#285
4. US-DHHS, Office of Civil Rights (OCR). Summary of HIPAA Privacy Rule, May 2003. http://www.hhs.gov/ocr/privacy/hipaa/understanding/summary/index.html
5. National Human Subject Protection Advisory Committee (NHRPAC) Recommendations on Public Use Data Files, January 2002. http://www.hhs.gov/ohrp/archive/nhrpac/documents/dataltr.pdf

Appendix B-1

Does Your Research Involving Secondary Data, Documents or Biological Specimens (*collected by a third party and provided to you for research) Require Review by the Cornell IRB Office?**



*Contact the Cornell Office of Sponsored Programs (www.ovpr.cornell.edu/osp) if acquiring the data requires a Data Use Agreement or a Materials Transfer Agreement between the provider and recipient.

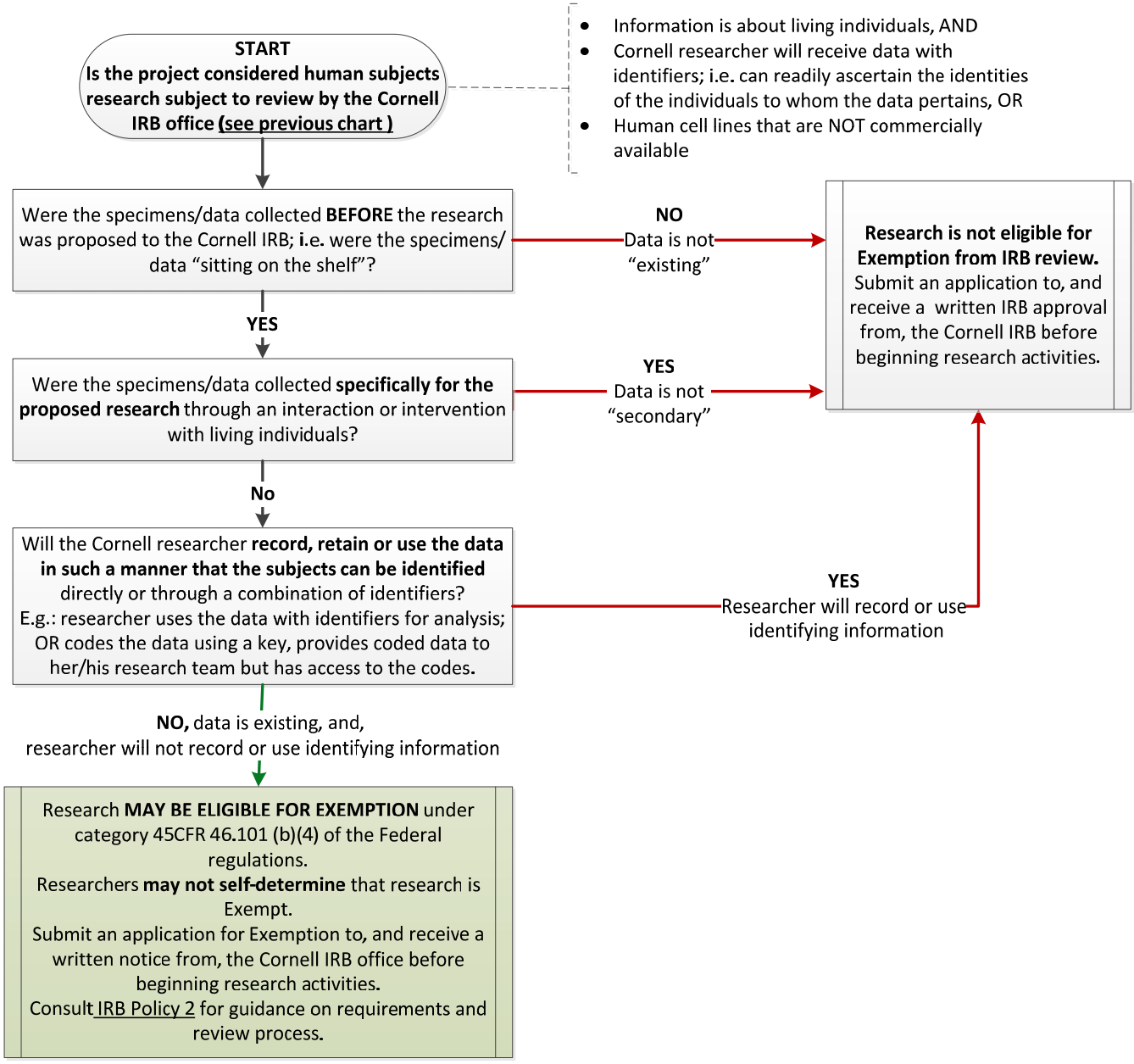
Reference:

"Research Involving Private Information or Biological Specimens Flowchart", National Institute of Health (NIH), January 2006, https://grants.nih.gov/grants/policy/hs/PrivateInfoOrBioSpecimensDecisionChart.pdf

Appendix B-2

Is Your Human Subjects Research Project Involving Secondary or Existing Data, Documents or Biological Specimens Eligible for Exemption from IRB Review?

Refer to Cornell IRB Policy #2* for application and review requirements for Exempt projects



- ***Cornell IRB Policies**: www/irb/cornell.edu/policy
- **Contact the Cornell OSP** if acquiring the data requires signing an agreement with terms and conditions for use, or a Materials Transfer Agreement. IRB approval or Exemption will be contingent upon the execution of agreement.

Appendix C

Typical Confidentiality Statement Components

A typical data sharing agreement includes the following sections (including the highlighted elements for confidentiality of de-identified data without the key).

Period of agreement: Clearly define a) when the provider will give the data to the receiver; b) how long the receiver will be able to use the data; and c) what will happen to the data once the agreement expires (e.g. deleted from hard drives, shredded, burned, etc.).

Intended use of the data: State as specifically as possible a) how the receiver will use the data; b) the studies to be performed or questions to be asked; and c) whether the researchers may use the data to explore additional research questions without approval or consent by the data provider.

Constraints on use of the data: List any restrictions on a) how the data or data findings can be used; b) whether the receiver can share, publish or disseminate data findings and reports without the approval or review of the provider; c) when the receiver generates a report based on the data, to whom does the report belong ; and d) whether the receiver is permitted to share, sell or distribute any part of the database.

Data confidentiality: State a) that the provider has removed all individual identifiers from the data; b) that the key to the coded data is not being shared with the receiver; c) describe all safeguards in place to ensure that individuals are not identified indirectly (e.g. through a combination of specific characteristics and location); and d) describe all safeguards in place to prevent sensitive information (e.g., salaries, exam results) from becoming public.

Data security: Describe the methods that the receiver must use to maintain data security, including a) locked cabinets and rooms, and password protection of electronic copies of data; b) which personnel will have access to data; c) the types of password protections and encryption to be used; and d) disposition of the data after the data-sharing period ends.

Methods of data-sharing: Identify the way in which data will be transferred from the provider to the receiver including a) whether the data will be transferred physically or electronically; b) how a secure electronic connection will be guaranteed; and c) if, and how, the data will be encrypted before being transferred.

Financial costs of data-sharing: Clarify who will cover the monetary costs of sharing the data.

References

Adapted from The Capacity Project. Available at: http://www.ihris.org/toolkit/tools/data-sharing.html.