Long-term Longitudinal Study Use Case
Summary Description

Research centers based at universities or government agencies often conduct studies that produce longitudinal microdata (individual-level rather than aggregate data) describing health, socioeconomic, and behavioral characteristics of human subjects. The data are collected by field investigators using questionnaires, interviews, and direct measurements and are stored in tabular datasets hosted by a research center or another institutional repository. Field investigators generally collect data every one to five years, and the research centers make these datasets available in waves. Researchers follow participants over time, and, particularly for health studies conducted over the course of many decades, they often include the children and grandchildren of the original participants in the study.

The data are generally disseminated in microdata form via a repository hosted by the research center or another institution. The datasets are relatively small, with 10,000 - 100,000 observations, and a total size of under 5 GB. When hosted by a large repository, web-based data analysis and visualization tools are often provided. These tools range from simple crosstabulation and frequency tables to regression analysis.

Prior to collecting human subjects data, researchers must obtain signed consent agreements from participants. Some of the information collected by researchers may be highly sensitive; for example, the National Longitudinal Study of Adolescent Health collects data related to adolescents' mental health, suicide intentions/thoughts, medications, religiosity and spirituality, substance use/abuse, violence, delinquency, sexual behavior, and sexually transmitted infections.[1] Researchers may also request educational data, such as transcripts, from the schools attended by the human subjects.[2] Therefore, these datasets may contain information protected by the Family Educational Rights and Privacy Act (FERPA). Researchers may also collect medical records from hospitals and physician offices covered by the Health Insurance Portability and Accountability Act (HIPAA).

To protect the privacy of human subjects, researchers or repositories take a number of steps. Generally, they remove information that directly or indirectly identifies an individual, including geolocation information, and require researchers who download the data to agree to a data use agreement. These agreements typically require data users to use the data only for research purposes and prohibit them from attempting to identify individuals or link the data with another dataset that could identify individuals. They may also partially or completely prohibit researchers from transferring the data to third parties.

Researchers may make available two separate datasets: public-use and restricted-use, where the public-use dataset is often made available for download to individuals who provide a name and email address. A public-use dataset may

[1] Add Health, Design Facts at a Glance, http://www.cpc.unc.edu/projects/addhealth/design/designfacts.
[2] *Id.*

contain only a subset of the total number of observations or attributes collected, while a restricted-use dataset may contain all the data. To gain access to a restricted-use dataset, a researcher may need to demonstrate that she holds a full-time, permanent, doctoral-level faculty appointment and/or her data protection plan has been approved by a human subjects institutional review board. Other studies, such as the Framingham Heart Study, make their data available only by application.[3]

Longitudinal studies can be extremely long lived and collect rich qualitative information. These characteristics create a number of challenges for managing privacy over time. First, the value of the studies lies in part in their ability to link a set of behaviors and changes to each individual over time, but this tends to make the combination of observable characteristics associated with each subject (quasi-identifiers) unique, and thus potentially identifiable. Second, the samples collected by these studies are rich enough to support analysis methods and research questions not originally envisioned in the study design. Third, analytics methods, methods for identifying privacy risks, and privacy regulations are all likely to change over the lifetime of the data collection.

Examples

1. Framingham Heart Study (FHS), an epidemiologic study begun in Framingham, Massachusetts, in 1948 with 5,209 men and women. Every two years, the FHS conducts physical and psychological examinations of and collects blood and tissue samples from study participants. Since the start of the study, the FHS has studied three generations of participants resulting in biological specimens and data from nearly 15,000 participants.

References:

- Framingham Heart Study, http://www.framinghamheartstudy.org

2. National Longitudinal Study of Adolescent Health (Add Health), a longitudinal study of a nationally representative sample of over 90,000 adolescents (and over 10,000 of their parents) in grades 7-12 in the United States during the 1994-95 school year. The Add Health cohort has been followed into young adulthood with four in-home interviews, the most recent in 2008, when the sample was aged 24-32. Add Health combines longitudinal survey data on respondents' social, economic, psychological and physical well-being with contextual data on the family, neighborhood, community, school, friendships, peer groups, and romantic relationships, providing unique opportunities to study how social environments and behaviors in adolescence are linked to health and achievement outcomes in young adulthood.

[3] Framingham Heart Study, Research Application Overview, Review Process, and Procedures, http://www.framinghamheartstudy.org/research/review.html.

References:

- Add Health Study, http://www.cpc.unc.edu/projects/addhealth

3. Panel Study of Income Dynamics (PSID), a national study of socioeconomics and health over lifetimes and across generations. The study began in 1968 with a nationally representative sample of over 18,000 individuals living in 5,000 families in the United States. Information on these individuals has been collected continuously, including data covering employment, income, wealth, expenditures, health, marriage, childbearing, child development, philanthropy, education, and numerous other topics.

References:

- Panel Study of Income Dynamics, http://psidonline.isr.umich.edu