ASHG
American Society of Human Genetics

Success Stories
in Human Genetics and Genomics Research

DATA SHARING TO ADVANCE GENETIC AND GENOMIC KNOWLEDGE

Data sharing has been common practice in genetics and genomics research for decades, accelerating biomedical discoveries and improving human health. Sustained funding to the National Institutes of Health (NIH) will ensure that genetic and genomic data remain accessible to researchers in ways that advance science and also protect the privacy of participants and minimize the stigmatization of groups of people.

Large-Scale Data is Transformative

Researchers use biobanks – large collections of biological samples and health information – and other research resources with shared genetic and genomic data from large numbers of diverse research participants to investigate new scientific questions. Open data empowers investigators with the ability to pool data, effectively increasing the sample size for robust studies and promoting reproducible science. Reusing existing data this way helps scientists learn more about the health of individuals and populations and drives medical innovations.

Improving Genetic and Genomic Knowledge Through Federally Funded Resources

NIH funding for human genetics and genomics has directly unleashed the power of precision medicine to improve health and well-being. These are a few examples of federally funded resources available to approved researchers:

What is Data Sharing?

Data sharing is the practice of making de-identified sociodemographic, genomic, and medical record data used for research available to other investigators. Sharing research findings benefits the scientific community, fosters collaboration, and increases transparency with the public. Since science progresses by building upon the work of others and innovates from prior discoveries, ethical data sharing accelerates knowledge.[1]

- **The *All of Us* Research Program.**[3] The mission of this NIH-funded research program is to enroll a diverse group of at least 1 million people from across the U.S. to share health information and biological samples. These data empower researchers to study individual differences in lifestyle, socioeconomic factors, environment, and biological characteristics to advance precision diagnosis, prevention, and treatment. Many of the participants are from groups that have been historically underrepresented in biomedical research, and the program aims to make the research results accessible to participants. To use this resource, researchers must go through training on patient privacy protections and ethical use of sensitive data. There are currently over 5,000 research projects leveraging *All of Us* resources, allowing researchers to define genetic risks better for conditions like heart disease, high blood pressure, and diabetes, underscoring the value of data sharing for health research.

- **Million Veterans Program (MVP).**[4] This ongoing study in the Department of Veterans Affairs Healthcare System is designed to look at how genes, lifestyle, military experiences, and environmental exposures influence health and disease among veterans. It is the nation's largest genomic biorepository of Veteran data and is one of the most diverse cohorts of any genetic research program in the world. Only approved

researchers may access the data, and participant security and confidentiality are protected through a variety of methods. Data from this program has led to new findings about anxiety, post-traumatic stress disorder, heart disease, kidney disease, cancer, and other health conditions.

- **The database of Genotypes and Phenotypes (dbGaP).**[5,6] This NIH-sponsored database was created to archive, curate, and distribute data from studies investigating the interaction of genotype and phenotype (the observable traits that people express resulting from the interaction between their genes and the environment). To access individual-level data, researchers must apply for controlled access by stating their research objectives and demonstrating their ability to keep the data safe. In the past five years there were over 400,000 approved data requests from around the world, with over 2,000 publications citing dbGaP each year, enabling better understanding of the genetic underpinning of diseases and traits.

Big Data, Big Responsibilities

To continue to realize the benefits of data sharing, it is essential to maintain protections for research participants' privacy. There are laws already in place protecting the privacy of participants in research funded by NIH and other federal institutions. Most genetics and genomics research is federally funded, and is thus subject to these laws.

Specific biobanks or data repositories may also have their own regulations to ensure that data are shared responsibly. For example, to access data from the *All of Us* program, researchers must complete training on data transparency, privacy, and ethical uses of genetic data. But even at the highest level of access, individual information is anonymized, meaning it cannot be legally linked back to an individual. Also, privacy-preserving machine learning techniques such as federated learning are novel strategies for secure data sharing among investigators.[7]

Genomic Data Sharing Policy

The NIH maintains many repositories for genomic data and expects the institutions and researchers it funds to abide by a set of regulations called the Genomic Data Sharing Policy. These expectations are meant to ensure that genomic research data are shared confidentially and responsibly with approved investigators.[2]

How Can Congress Support Research?

Data sharing is accelerating the rate of biomedical discoveries and innovations that impact people's lives and health. To maximize these benefits, policies must encourage broad public participation and promote privacy protections.[8] NIH funding is critical to ensure that the public is engaged in research and their data are shared responsibly.

In addition, researchers need robust and predictable NIH funding. Congress can continue to support the large-scale data sets necessary for genetic and genomic research through sustained NIH funding.