GUIDANCE: Best Practice on Depositing Data into Large Databases

Increasingly the NIH and other federal funding sources insist that data, especially genomic data, generated in research are put into large nationally accessible databases. If a repository is funded by NIH (such as dbGAP or dbSNP), there are very particular specifications that must be met before researchers can deposit data. Per the requirements of the NIH, the IRB needs to be able to certify certain things. In order to do so, researchers need to have specific information in the protocol regarding the data sharing plan and language in the consent form describing what the researcher is doing with the data, the risks unique to depositing large amounts of genetic data, and what it means to have one's data in such a database. While this article highlights work funded by NIH and other federal sources, the IRB supports a single standard approach and regardless of funding, suggests all projects meet the criteria listed below if depositing data into repositories. Links to the NIH Policy are at the bottom of this article.

If researchers plan to deposit data into large databases, a clear statement regarding deposits must be in the consent form and best practice is to include an option for subjects to agree/not agree to this use of their research data in the consent form.

Please read this information carefully and keep the suggested language available when crafting consent forms. It is recommended that you modify your current project and consent forms if your vision is/was to share data and you don't currently have that process approved in your project.

Data sharing plans include data type, data repositories to which data will be submitted, appropriate uses of the data and limitations on the future use or an exception to submission which explains why the Institutional Certification criteria cannot be met and must describe an alternative mechanism for data sharing. Please see the NIH Guidance for Investigators in Developing Genomic Data Sharing Plans for more information.

If you are planning on depositing data into a large-scale repository, NIH requires the IRB must assure certain criteria are met:

1. The protocol for the collection of data is consistent with 45 CFR 46;
2. Data submission and subsequent data sharing for research purposes are consistent with the informed consent of study participants from whom the data were obtained;
3. Consideration was given to risks to individual participants and their families associated with data submitted to NIH-designated data repositories and subsequent sharing;
4. To the extent relevant and possible, consideration was given to risks to groups or populations associated with submitting data to NIH-data repositories and subsequent sharing; and
5. The investigator's plan for de-identifying datasets is consistent with the standards outlined in the GDS (Genomic Data Sharing) Policy.

In order for the IRB to make the assurance, it needs to review the protocol, data sharing plan details, and informed consent document.

The protocol needs to include information about the data sharing, future use and deposits to repositories, in addition to statements about the potential privacy risks to the subjects and their families (as well as to their kin group or population, if a specific population or group is being targeted or investigated).

The informed consent document must include detailed information about future use including what data will be shared, with whom and for what purposes. In addition to detailed future use language, the following language must be used:

IRB Approved Language for Consent Form(s):

Under "Possible Risks and Discomforts"

Group Risks

Although your name will not be given to researchers, basic information such as your race, ethnic group, and sex may be shared. This information helps researchers learn whether the factors that lead to health problems are the same in different groups of people. It is possible that such findings could one day help people of the same race, ethnic group, or sex as you. However, they could also be used to support harmful stereotypes or discrimination.

Privacy Risks

Your name and other information that could directly identify you (such as an address, date of birth, or social security number) will never be placed into a database [if true]. However, because your genetic information is unique to you, there is a small chance that someone could trace it back to you. The risk of this happening is very small, but may grow in the future. Since the database contains genetic information, a break in security may also pose a potential risk to blood relatives as well as yourself. For example, it could be used to make it harder for you (or a relative) to get or keep a job or insurance. If your private information was misused, it is possible you would experience other harms, such as stress, anxiety, stigmatization, or embarrassment from revealing information about your family relationships, ethnic heritage, or health conditions.

Insurance Risks

There is a Federal law called the Genetic Information Nondiscrimination Act (GINA). This law makes it illegal for health insurance companies, group health plans, and most employers of over 15 people to discriminate against you based on your genetic information. However, it does not protect you against discrimination by companies that sell life insurance, disability insurance, or long-term care insurance.

For more information on the GDS policy, data use limitations, writing data sharing plans, and more visit https://sharing.nih.gov/genomic-data-sharing-policy/about-genomic-data-sharing/gds-policy-overview