

**Can a PI delegate data loading / submission to another individual on his/her research team?**

Absolutely. The ability to designate another individual to submit is built into our submission portal. In the submission portal you have a way to invite people -- you can add submitters, so when somebody is originally invited and that is either the PI or what we call the PI uploader/assistant in the registration system, but you could add other people.

**Are there links between dbGaP and ClinicalTrials.gov?**

We have just started to establish links between dbGaP and ClinicalTrials.gov. We do not have a lot of these links yet, but we do have a number of them. Recently we added a way to reference a ClinicalTrials.gov number in a submission. Usually the ClinicalTrials record is established before the dbGaP record; so dbGaP is in a better position to get the link from the submitter.

**Are there any required variables/phenotypes that need to be included, aside from subject id, of course?**

Each of those core tables that I mentioned in the context of our submission guide has required elements. For example in the subject consent table, the subject ID and the consent are required. That is absolutely required. There are optional fields where you can reference other samples and aliases of the sample name itself. In some cases additional information is required. For example if you submit a pedigree file, you have to provide all five columns, which family, person, maternal id, paternal id, sex. Read the submission guide for each file that you are creating. It describes what is optional what is not. For extramural, studies, there is usually an expectation from the funding Institute for a minimal set of phenotypes included in the study. Specifically, that the main variable (e.g., heart disease) and co-variates (e.g., age, weight, sex) used in the analysis are submitted to dbGaP so that other people can reproduce the information in your publication. Intramural Investigators should discuss with their GPA. The goal is to include the data that would be required for another researcher to be able to reproduce the published analysis.

**Are there any standards implemented for data submission like MINSEQE for Next Gen data or MIAME for microarray data or TREND for Clinical, Non-randomized trials?**

The genotype submission guide describes the formats most commonly submitted and includes all types of molecular data, but it does not list them all. With that said, we will take almost any standard form of data. The bulk of the data coming in now is in .vcf, .maf or PLINK. We do have MIAME standard expression array data as well.

**Is there a way to update publications on the data after submission?**

Absolutely. We can update publications in between versions. You do this without update the data, and we would not increment the version of the study. The phenotype curator you were assigned initially when you submitted your data would normally update the publication list. You would simply contact that person with the new publication information and they can update the page with in a day usually.

**Do we have to submit the raw data to SRA, VCF files to dbGAP on the same NGS dataset?**

You don't necessarily have to submit both, but I think it is highly desirable. The SRA data is not as useful to as many people as the VCF file just because of the sheer size and computation involved in variant calling. We at dbGaP have to refer to the program officers and GPAs on what their expectation is. For intramural scientists, I think that is something you should discuss with your GPA. From a data sharing point of view, I think submission of both the variant calls (VCF) to dbGaP and the reads (SRA) would be best.

**The invitations to submit seem to expire quickly. Invitations expire quickly. Why?**

Yes they do expire quickly. The invitation to submit should be accepted within one week. If not, we have to reissue it. We don't want the invitation to be out in the open for too long. We are following the model that was set up by the manuscript submission process for PubMed Central. It's simple to accept the invitation. Once you accept, it does not expire. You can log on any time after that to upload data.

**Please describe the release schedule for uploaded data, and touch on "exchange areas" and how those are used, and whether that data is also released.**

I think the GDS policy is that data are released six months after the data are produced. It has been somewhat open to interpretation as to when exactly is the moment when the data are produced. For instance, is it when it comes off the machine? My understanding is that most people are interpreting that the 6 months starts when the data is QC 'd by the group producing the data. The QC process does take some time, and is not really included in the six months. There is no embargo on the GDS policy. Previously under the GWAS policy, we would release the data, but it would be embargoed for some time before anybody could publish on it. With the GDS policy, there is no embargo, but there is a protected set of time for the data producing researcher to process and publish on his/her own data.

Exchange Areas refer to a pre-release version of a dbGaP study and are intended to facilitate large multicenter project's QC effort. The data can be submitted to dbGaP and SRA as a sort-of storage and exchange area. Disparate QC teams can access this area using a request process similar to the normal Data Access Request. If you have a study that you think can benefit from an Exchange Area, please contact me, Mike Feolo, and your GPA to discuss this before registering your study.

**What's the detailed dbGaP policy on cell line data? Note that currently there is cell line data in dbGaP. If I'm submitting data generated using a commercially available cell line and ultimately want the raw sequence reads to be uploaded to GEO/SRA. Also, I want unrestricted access for the data. Do I need my studies registered in dbGaP? My understanding is that dbGaP only allows restricted access the data I submit. Is this correct? These were HEK cell lines derived in the 70s.**

Take a look at the GDS policy site (<https://gds.nih.gov/>). It depends on when the cell lines were collected. You have to pay attention to that, and I think if you work with your GPA, you will be able to figure out what is appropriate. I just wanted to show you that if you go to the

Researchers section of the GDS site (<https://gds.nih.gov/06researchers1.html>), you will see in the data institutional certifications are actually set up so there for different time frames. There are also separate sections for extramural investigators and intramural investigators.

**Can a PI amend a study by after first uploading subject level data, for example, if a PI enrolled a new cohort in subsequent years and wanted to add these data to the original dbGaP submission?**

Yes, this is what I meant by a study can be versioned. This is quite common, for example The Framingham Heart study has something like 27 version thus far. The study version changes whenever additional data is added, redacted or updated. We expect that a study will change versions no more often than once a quarter. This is to provide a stable version for end-users and for pragmatic reasons in terms of the dbGaP team's time.