

Genomic Data Submission: Investigator Checklist

According to the [NIH Genomic Data Sharing \(GDS\) Policy](#), investigators are required to deposit all read-level human genomic data in dbGaP. Non-PHI data can be published so it's publicly available via SPARC. Thus studies generating human sequence data that produce both PHI and non-PHI data follow the two-repository workflow illustrated in Figure 1 below. This [SPARC/dbGaP Quick-Reference](#) gives an overview of what file types should go into dbGaP vs. SPARC. Please contact your NIH Program Officer or Genomic Program Administrator (GPA) if you have questions regarding specific dbGaP submission file type requirements.

Overview of Process

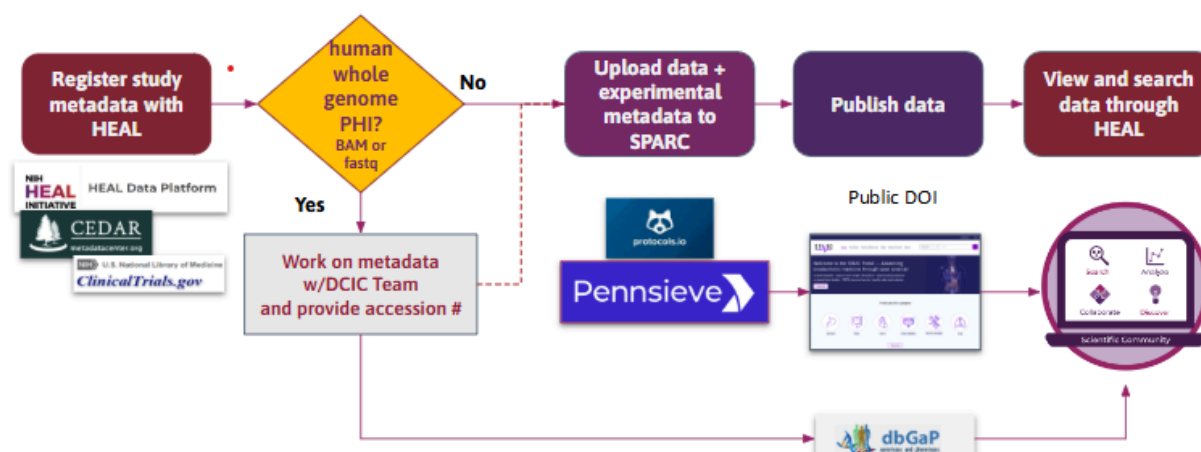


Figure 1. SPARC process for HEAL data that includes dataset submission to dbGaP.

Subject Tracking

Submitting multiple datasets from the same subjects (especially across repositories) requires careful tracking to ensure the full dataset can be both fully understood and reconstructed correctly. Designate a point person to maintain your project's master subject ID spreadsheet so IDs are coordinated across all experiments, collaborators, and datasets (an example of what this might look like is illustrated in Figure 2). Additional details can be found in [Linking subjects across datasets](#).

Your Lab Master Subject IDs	Record repository and Dataset DOI or Accession # of each dataset ->	SPARC Dataset 1 Subject IDs	SPARC Dataset 2 Subject IDs	dbGaP Dataset 1 Subject IDs
		https://doi.org/10.26275/vpx0-bk5y	TBD (Pending)	phs004229.v1.p1
1		sub-001		sub-001
2		sub-002		sub-002
3	Record subject IDs used in a study ->	sub-003	sub-003	sub-003
4		sub-004		sub-004
5		sub-005	sub-005	

6		sub-006		
7			sub-007	sub-007
8			sub-008	
9			sub-009	sub-009

Figure 2. Example of a Master subject ID spreadsheet. Subject IDs **must be identical** across SPARC and dbGaP submissions.

Step 1 — SPARC Onboarding & ID Coordination

- Contact the SPARC Curation Team (curation@sparc.science) to discuss your submission and workflow.
- Determine whether your data contains files that should be submitted to dbGaP (see [SPARC/dbGaP Quick-Reference](#)).
- Share your “Master subject ID spreadsheet” with the curation team to facilitate tracking subjects across submissions.
- Obtain access to the workspace associated with your SPARC study on Pennsieve - contact Sam Kessler (samkessler@sparc.science) or the Pennsieve team (support@pennsieve.io) to create an account and obtain access.
- Prepare SPARC [metadata files](#) following instructions from The Curation Team (subjects.xlsx, samples.xlsx, dataset_description.xlsx, manifest.xlsx).
- Create a dataset placeholder on the Pennsieve platform and reserve a DOI. Share this dataset with the Curation Team.

Step 2 — dbGaP Registration

- Register your study in dbGaP and obtain a dbGaP accession number (*phs#*).

Step 3 — SPARC Data Upload & dbGaP File Conversion

- Enter the dbGaP study accession number in the master subject /sample ID spreadsheet for each submission (see *second row in Figure 2 above*).
- Record the phs# in your SPARC *dataset_description.xlsx* file (Related Identifier field).

32	Related protocol, paper, dataset, etc.		
33	Identifier description	dbGaP study ID	
34	Relation type	IsDerivedFrom	
35	Identifier	https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs00XXXX.v1.p1	
36	Identifier type	URL	

Figure 3. Example of dbGaP information in a SPARC *dataset_description.xlsx* file.

- Notify the SPARC Curation team once you have added the phs# in your *dataset_description.xlsx*.
- Complete and upload all required metadata and non-genomic experimental data to [SPARC / Pennsieve](#).
- Convert SPARC metadata files using the Pennsieve tool on the platform to generate your dbGaP submission files.
- Ensure Subject and Sample IDs match across submission metadata.

Step 4 — dbGaP Sequence File Submission

- Submit human sequence files to dbGaP and metadata files generated in Step 3. Include documentation linking this submission back to the corresponding SPARC dataset and contact the Curation Team for current documentation guidance.
- Await dbGaP data review, dbGaP representatives will notify you when the review is complete.

Step 5 — SPARC Dataset Publication (This DOES NOT rely on your dbGaP submission and can be done before, in parallel with, or after your dbGaP submission)

- Request to publish your SPARC dataset on Pennsieve.
- Await SPARC dataset review - you will be notified when the review is complete.

Step 6 — Release & Discoverability

- Once dbGaP review is complete and the dataset is published, the controlled data will be accessible via [dbGaP Authorized Access](#).
- Once SPARC review is complete, and the dataset is published, the data will be open access and available via the [SPARC Portal](#).

Key Contacts

Name	Role	Contact
SPARC Curation Team	SPARC data organization and publication assistance	curation@sparc.science
Pennsieve Help Desk	General data upload and SPARC → dbGaP Converter Tool (Pennsieve)	support@pennsieve.io
Your NIH GPA	Genomic Program Administrator — contact your NIH Program Officer	Via the study portal
dbGaP Help Desk	Submission Portal questions	dbgap-sp-help@ncbi.nlm.nih.gov

Useful Links

- [NIH GDS Policy](#)
- [Repository Allocation Quick-Reference Table](#)
- [SPARC / Pennsieve](#)
- [SPARC Submission Guide & Templates](#)
- [SPARC to dbGaP Conversion Guide](#)
- [dbGaP Submission Guide & Templates](#)
- [dbGaP FTP Template Downloads](#)
- [GapTools — pre-submission QC validator](#)