Submission of WHI Data to dbGaP - Guidelines

WHI policy and guidelines related to the submission of genetic data to dbGaP by Ancillary Study Investigators

Background

WHI Ancillary studies may be required to or choose to submit their data to dbGaP (database of Genotypes and Phenotypes) a central data repository at the National Center for Biotechnology information (NCBI). Submission is a condition of the funding agency in many cases. WHI has specific limitations on the data that can be submitted based upon WHI participant consent. Therefore, Ancillary Study investigators submitting genetic data should work with the WHI-CCC and follow the steps outlined below to ensure the data is submitted correctly. Please contact the WHI-CCC with any questions or help using these guidelines.

I. WHI "top level" parent study (phs000200) and Ancillary Study Substudies

1) WHI is a "top level" parent study on dbGaP (phs000200).
   a) It contains a study description, participant numbers, available variables, and datasets (molecular and other).
   b) Most ancillary studies are "substudies" to the parent WHI study.
   c) The WHI-CCC submits updated phenotype data to phs000200 annually for all WHI participants who have the appropriate consent for dbGaP.
   d) The WHI-CCC maintains the Subject Consent for all WHI dbGaP participants.
   e) Substudy pages are added by the AS Principal Investigator as needed.

II. Other WHI data on dbGaP

1) Some WHI Ancillary Studies are stand-alone "top level" studies, or substudies to other "top level" studies (e.g., consortium studies involving one or more other cohorts).
   a) It was discovered in 2016 that the proper consent groups and Data Use Limitations (DULs) were not used for some of these studies, a potential violation of WHI informed consent.
      i) Re-registering data as a substudy to WHI was logistically complex and a waiver was granted at the request of the funding agency to allow the study to have their own page separate from WHI.
      ii) Investigators are required to recertify the data with the correct consent groups and DULs.
         (1) Each investigator will need to work with NCI or the relevant institute separately.
         (2) Ideally the consent group/DUL updates will occur with a regularly planned submission of updated Ancillary Study's dbGaP data, if possible.

III. Future WHI Ancillary Study data on dbGaP

1) Per WHI Ancillary Study Policy, Ancillary Studies required or desiring to submit data to dbGaP should register and submit data as a substudy to WHI.
   a) It is WHI's obligation to live up to the informed consent.
   b) This ensures that the correct consent groups and DULs are applied.
   c) Updates are easier to handle, e.g., withdrawal of participant consent, revisions to DULs and acknowledgements.
   d) Requests for data are streamlined with one phs accession number covering all phenotype and genotype data.
   e) WHI phenotype data updates are done by WHI-CCC (exception: some complex non-molecular phenotype data might need to be submitted by the study PI).
   f) Note: Data must be submitted to the CCC per usual policy, in addition to dbGaP.
2) CCC will provide a "WHI certificate", to be used by the submitting institution for preparation of the official Institutional Certification (IC).
3) Requests to submit as a standalone study or substudy to another top level study, per a condition of the funding agency (e.g., consortium studies with several cohorts):

a) PI confirms requirement with PD/PO of funding award.
b) PI submits request for an exception to the WHI Project Office.
c) PI is responsible for ensuring correct consent groups and DULs are applied (guidelines provided in this document).
d) PI provides CCC with the phs number of the study when fully registered.
    i) CCC keeps a record of all WHI data posted to dbGaP and corresponding phs numbers.

IV. Procedures and Guidelines for submitting WHI data to dbGaP

1) Limitations of WHI data
    a) There are four distinct groups of WHI data and their restrictions, per Fred Hutch IRB (late 2009).
        (1) Those who signed the supplemental use consent allowing their data to be used in genetic studies (72.7%).
            (a) may be posted for commercial and noncommercial use
        (2) Those who refused to sign the supplemental use consent (11.4%).
            (a) No individual level data may be posted
        (3) and (4) Those who did not respond to the supplemental use consent (8.6%), or those who were deceased at the time of the supplemental use consent (7.2%).
            (a) Individual level data may be posted for noncommercial use
    b) Study investigators receive an accounting of how many participants fall into each of the above categories after selection is complete. When it is known in advance that all study data must be submitted to dbGaP, the CCC will not provide specimen for participants who are not dbGaP eligible. If it is not a requirement that all data be submitted to dbGaP, the CCC will identify those who are eligible and those who are not eligible.

2) Institutional Certificate (IC) and dbGaP Consent Groups and Genomic Summary Results (GSR)
    a) As the depositor of the data, it is the investigator's responsibility to obtain the IC.
        i) Due at JIT for those studies that provided a genomic data sharing plan in the NIH grant application, per the implementation of the NIH Genomic Data Sharing Policy.
            (1) Check with your institution's IRB for the procedure to obtain the IC.
                (a) This will involve review of WHI Consent forms, available on the WHI website.
                (b) DbGaP template ICs are recommended: https://gds.nih.gov/Institutional_Certifications.html
                    (i) Template for WHI data is "For Studies Using Data Generated from Cell Lines Created or Clinical Specimens Collected BEFORE January 25, 2015" – – "that have consent"
                    (ii) WHI consent group titles and modifiers are:
                        1. Health/Medical/Biomedical-IRB (HMB-IRB)
                        2. Health/Medical/Biomedical-IRB-Non-profit-use (HMB-IRB-NPU)
                            a. Note: These consent group names were updated with WHI version 9 release on dbGaP (" Consent group 1 (c1) is changed from General Research Use (GRU) to Health/Medical/Biomedical (IRB) (HMB-IRB). Consent group 2 (c2) is changed from Non-Profit Use Only (NPU) to Health/Medical/Biomedical (IRB, NPU) (HMB-IRB-NPU"). This was an effort on the part of all NIH Data Access Committees (DACs) to standardize the consent language across all dbGaP studies. In making these changes, the DACs hope to avoid future confusion on the part of the applicants, and cut down on the number of Data Access Requests (DARs) which propose to use the datasets in a manner which is not permitted.
                    (iii) As of May 2019, Genomic Summary Results (GSR) should be marked on the IC as 'controlled access' (sensitive designation), with explanation:

"Genomic Summary Results should be maintained under controlled access as some of the data includes relatively small minority populations that are vulnerable and have been stigmatized in the past, as well as some sensitive phenotypes."

(c) NOTE: Ancillary Study IRB cannot override the WHI parent consent group categories/DULs.

3) Data Use Limitations (DULs)

a) Current WHI DULs (automatically applied to WHI substudies):
"All requestors and data users must provide a local Institutional Review Board (IRB) letter from their respective institution approving conduct of the proposed research project. Only full or expedited approvals will be accepted. There are two consent groups for this study:
a) Health/Medical/Biomedical – IRB, and b) Health/Medical/Biomedical – IRB for use by not-for-profit organizations only.

Use of the WHI data to conduct non-genetic research is prohibited. WHI data may not be used to investigate individual pedigree structures, individual participant genotypes, perceptions of racial/ethnic identity, variables that could be considered as stigmatizing an individual or group, or issues such as non-maternity. Users acknowledge that American Indian participants were self-identified and were primarily dwelling in urban areas and are not representative of the diverse American Indian population across the United States. Users agree not to use the data to infer tribal status or affiliation.

i) American Indian provision was also added 1/2017 in the Data Prep and Data Dictionary (as a Usage Note) for Form 2 and Form 41. It will be in the WHI Investigators datasets as part of the 6/2017 data release.

ii) Two lines were taken out in 2017:

(1) "The IRB letter must explicitly reference the use of WHI data. "Origin of this statement unknown.

(2) "All research must be related to the etiology and prevention of morbidity and mortality of post-menopausal women, consistent with the WHI model informed consent documents." Technically, this statement is covered by the HMB designation.

b) Some DULs are automatic per the consent groups and modifiers listed on the IC: https://gds.nih.gov/pdf/standard_data_use_limitations.pdf

c) Other DULs are modified by informed consent and IRB review, and NHLBI.

i) Revisions may be suggested by NHLBI based on experiences by one or more studies.

(1) American Indian language added by the PAGE study.

(2) Waiver of IRB not permitted. Many IRBs consider accessing data as non-human subjects research, but given the depth of phenotypes for WHI, full or expedited review was added in.

(3) Some language may be redundant with the boilerplate of the DUC and the Code of Conduct.

(4) Revisions to DULs apply automatically to all WHI substudies.

(5) Stand-alone studies will be notified of any revisions and will need to revise their studies separately.

d) DULs must be part of the IC

i) Existing WHI DULs automatically apply to WHI substudies, but must still be listed in the IC.

ii) Space not provided in template IC.

(1) Print the form, cut and paste DULs in before the page 3 signatures and rescan and sign to have a complete form.

(2) Attach as an addendum, as long as it is clear that the signature covers the entire package.

e) DULs are also listed in the universal WHI Data Use Certification (DUC). Stand-alone studies may or may not list DULs in the DUC.

4) Registering data – Negotiation of Data Use Certification (DUC)
   a) As a Substudy to WHI
      i) Work with the NHLBI Genomics Program Administrator (GPA) to register under the WHI substudy.
         (1) Current Genomic Program Administrator (GPA; December 2016) for WHI is: Alex Runko / Sue Chen @ nhlbigeneticdata@nhlbi.nih.gov
         (2) Note: GPA first point of contact for TOPMed is NHLBITOPMeddbGaP@nhlbi.nih.gov; goes to Rebecca Beer and Mollie Minear
      ii) Submit a Data Submission Information sheet specific to the study, along with the institutional certificate (IC).
         (1) Refer to main WHI study whenever possible (phs000200), so updates can be made in one place.
      iii) Data Use Certification (DUC): No effort is needed by AS PIs. Substudies to WHI automatically fall under the WHI universal Data Use Certification (DUC).
         (1) DUC may be updated (rare) by the NHLBI GPA in collaboration with WHI-CCC.
         (2) WHI DUC includes the DULs.

   b) As a stand-alone study (i.e., exceptions granted by NHLBI PO).
      i) Registration documents may differ per the NIH institute
      ii) Please refer to the main WHI study (phs000200) whenever possible in all paperwork so updates need to be made in one place only.
      iii) Stand-alone studies will need to negotiate the DUC with their GPA as part of the dbGaP registration process.
         (1) Note: NCI DUCs do not list DULs. Instead, DULs in its entirety is in the registration, and on the study page.
      iv) Some phenotype data is sensitive and cannot be uploaded to dbGaP.
5) Submission of files via the submission portal after study registration
   a) After registration is complete, an automatic submission invitation and instructions are sent to the substudy PI and, if specified, secondary contact/data submitter.
   b) For details of the dbGaP submission process, refer to https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/GetPdf.cgi?document_name=HowToSubmit.pdf and follow the links to the Submission Guide.
   c) [As of 1/2017) Contact the WHI dbGaP Curator (Anne Sturcke, kianga@ncbi.nlm.nih.gov) and the dbGaP Submission Specialist (Lora Ziyabari, ziyabarl@ncbi.nlm.nih.gov) for details. These dbGaP staff members have been with dbGaP since the beginning. They are very knowledgeable about WHI dbGaP submissions, and they are also very responsive and helpful.
   d) The required files for a WHI Sub-study will be:
      i) Study Config document
         (1) Refer to main WHI top level study as much as possible to avoid duplication and make updates easier to manage (e.g., WHI study description, WHI study timeline).
            (a) Study config: Standalone studies (exceptions) may wish to add language to say additional data is available under WHI study (phs000200).
      ii) Subject Consent file and Data Dictionary
         (1) For WHI Sub-studies, this will be a blank txt file. The data submitted by the sub-study will be automatically linked to the WHI Subject Consent file by dbGaP by the Subject ID. (The WHI-CCC is solely responsible for keeping the Subject Consent file up to date with any changes of participant consent status.)

(a) Note: For stand-alone studies that need to submit these files, make sure to use the correct consent group names HMB-IRB and HMB-IRB-NPU.

iii) Subject Sample Mapping file and Data Dictionary
(1) This file will contain the WHI dbGaP Subject ID and (if applicable) its link to the WHI dbGaP Sample ID. (See Study IDs, below.) If the dbGaP Sample ID is not being used for the data submission, these files will be blank txt files.

iv) Sample Attributes file and Data Dictionary

v) The Ancillary Study data files and associated Data Dictionaries

e) The substudy accession number is assigned after the study config document and other files are submitted via the dbGaP submission portal.

6) Study IDs: NIH policy is two-step de-identification for dbGaP data, therefore dbGaP IDs are created for your samples. *The data you submit to dbGaP should be linked only to the WHI dbGaP Sample or Subject ID*.
a) A 'Common ID to dbGaP Subject ID' to 'dbGaP Sample ID' link file will be provided by the CCC.
b) *The link between the dbGaP Sample or Subject ID and the WHI Common ID or Draw/Vial ID is never to be submitted to dbGaP. The WHI Common ID or Draw/Vial ID is never to be submitted to dbGaP.*
c) The WHI dbGaP Sample ID corresponds to the WHI Draw ID. The purpose of the dbGaP Sample ID is to enable dbGaP to tie specific samples attributes to samples of differing platforms, days, aliquots, etc. It may be submitted to dbGaP and included in the required Subject-Sample mapping file (dbGaP Subject ID to dbGaP Sample ID). However, if the information about the Sample, such as the days from randomization, sample type, etc, is included in another data file submitted to dbGaP, there is no requirement to submit data using the dbGaP Sample ID. However, if the WHI dbGaP Sample ID is available, it is recommended that it be used when submitting lab data to dbGaP.
d) If needed,it is acceptable to submit the Pull ID to dbGaP, but it is recommended to limit it to the first two sections (e.g., AS555-3).
e) WHI blind duplicate quality control samples do not have study data and are not to be submitted to dbGaP.

7) Acknowledgements
a) Acknowledgement of both the parent WHI and the specific study should be made.
i) A link in the universal WHI DUC refers to a master acknowledgements list.
(1) Uploaded into the submission system by the GPA.
(2) The WHI-CCC is the keeper of the Master acknowledgement list. Substudy PIs should contact the WHI-CCC to add their specific acknowledgements to the master list. The CCC will work with the GPA to update it within dbGaP.
(3) Standalone studies should ensure the correct acknowledgements for WHI are used.

V. Requesting WHI data

1) dbGaP data access is at the level of the WHI top level study. The requestor may use any or all of the top level and substudy data in keeping with their approved Research Use Statement.
a) The approval document should state that your **IRB** agrees that your proposed research use, as it appears in the request form, is consistent with your institution's policies. It should be clear from the study title in the approval letter that this letter is indeed associated with the data access request.

VI. Glossary of Terms (work in progress)

GPA = Genomic Program Administrator: Person responsible for working with investigators to facilitate study registration in dbGaP. GPAs also serve as the lead NIH IC point of contact for the NIH GDS Policy. Roster of IC Genomic Program Administrators

dbGaP Curator (phenotype and genotype) = person(s) responsible for QC of submitted phenotype and genotype data.

PD/PO = Project Director/Project Officer of the funding award. Provides oversight to dbGaP submission.

DUL = Data Use Limitation. Listed on template Institutional Certification (IC) per IRB review of the data to be posted by the submitting institution.

DUC = Data Use Certification.

DAC = Data Access Committee. Each institute has their own DAC and DAC chair, responsible for reviewing data access requests and ensuring the use of the data complies with the data use limitations.

DAR = Data Access Request. Application to request data from dbGaP.

GSR = Genomic Summary Results. summary statistics that may be computed by an NIH-designated data repository, to include systematically computed statistics such as, but not limited to: 1) allele frequency information (e.g., genotype counts and frequencies, or allele counts and frequencies); and 2) association information (e.g., effect size estimates and standard errors, and p-values). **In May 2019, WHI Leadership determined that GSR should be designated as sensitive and kept under controlled access.**

VII. Attachments/Appendix:

WHI template IC

WHI data use limitations as they appear on dbGaP

WHI DUC

WHI dbGaP acknowledgements

WHI IDs