Compiled Public Comments on NIH Request for Information: Processes for database of Genotypes and Phenotypes (dbGaP) Data Submission, Access, and Management (NOT-OD-17-044)

February 21, 2017 – April 7, 2017

Public Comments

1. Manish Gala, MD, Massachusetts General Hospital
2. Russ Altman, Stanford University
3. bahram namjou, cchmc
4. Linda Davis, NIH Human Genome Study
5. Ben Burkley, University of Florida
6. David Auble, University of Virginia
7. Elizabeth J Ampleford, Wake Forest School of medicine
8. wade berrettini, U. of Pennsylvania
9. Dongmei Sun, University of Alabama at Birmingham
10. Haixu Tang, Indiana University, Bloomington
11. Heping Zhang, Yale University
12. Jared Roach, ISB
13. David Kwiatkowski, Brigham and Women's Hospital
14. Ken Winters, Oregon Research Institute
15. Jessie Tenenbaum, Duke University
16. Cathy C. Laurie, University of Washington
17. Richard Shaw (& Matthew Young), Repositive Ltd.
18. N/A, Baylor College of Medicine
19. Edwin K. Silverman, Brigham and Women's Hospital
20. Ara Tahmassian, Harvard University
21. Lasse Folkersen, Technical University of Denmark
22. Phil Canakis, The Pennsylvania State Univ Hershey Med Ctr
23. FASEB, Federation of American Societies for Experimental Biology
24. Preston Campbell, CEO, Cystic Fibrosis Foundation
25. UDN Coordinating Center, Harvard Medical School/Undiagnosed Diseases Network (UDN)
26. Luis Serrano, Center for Genomic Regulation (CRG)
27. S.P.A. Drummond, PhD, President, Sleep Research Society
28. Li-San Wang, University of Pennsylvania
29. Liam Curren, Genomics plc
30. P. Pearl O'Rourke, Partners Health Care
31. Sheila M Reynolds, Institute for Systems Biology
32. Rajni Samavedam, Booz Allen Hamilton
33. Wayne Huggins, RTI International
34. Stacey Donnelly, The Broad Institute, Inc
35. Brandi Davis-Dusenbery, Seven Bridges Genomics
36. Mary Majumder for Project Team, Center for Medical Ethics and Health Policy, Baylor College of Medicine
37. Sarah Nelson, University of Washington
38. Mark Daly, Broad Institute
39. Heidi Rehm, Brigham & Women's Hospital
40. Joel Hirschhorn, Broad Institute, Boston Children's Hospital, Harvard Medical School
41. Angela Patterson, Mayo Clinic
42. Aras Eftekhari, Attain, LLC
43. Daniel MacArthur, Broad Institute of MIT and Harvard

Submission Date
02/21/2017
Name
Manish Gala, MD
Primary Purpose of dbGaP Use
Both Study Registration / Data Submission and Data Access / Download
What is your level of experience with dbGaP?
Experienced (used many times over the course of several years)
Role/Other Role
Scientific Researcher
Type of Organization/Other Type
Nonprofit Research Organization
Name of Organization
Massachusetts General Hospital

Information Requested

1. dbGaP Study Registration and Data Submission
The delay in which items are submitted and available to the scientific community is quite slow.

2. dbGaP Data Access Request (DAR) and Review
This process is absolutely abysmal in several DACs. Several DACs are able to efficiently process requests in 1-3 business days. However, the great outliarin efficiency has been the eDAC for NCI. While it's charter claims a review in 21 business days, more recently investigators have been waiting for 2 months to obtain permission. This wait time has been observed for renewals as well as new applications. For members of the cancer research community, this delay hurts scientific progress in a field particularly where genomics is heavily utilized. If wait times are this egregious, please consider allowing temporary access (favor beneficence) until such an infrastructure to handle the increasing numbers of requests has been made.

3. Policies for the Management and Use of dbGaP Data

- **Alternate controlled-access models**
- **Benefits and risks associated with the availability of genomic study summary statistics**
- **NIH is interested in public comments on benefits and risks of such reference use of dbGaP data for research participants, patients, and the scientific community**

4. General Comments. NIH welcomes general comments on any other topics with regard to dbGaP data submission, access, and management

Submission Date
02/21/2017
Name
Russ Altman
Primary Purpose of dbGaP Use
Data Access / Download
What is your level of experience with dbGaP?
Experienced (used many times over the course of several years)
Role/Other Role
Scientific Researcher
Type of Organization/Other Type
University
Name of Organization
Stanford University

Information Requested

1. dbGaP Study Registration and Data Submission
I have never submitted data.

2. dbGaP Data Access Request (DAR) and Review
I have come to consider dbGAP a "write only" database since data goes in, but rarely comes out. Of course, this is not literally true, but that is the impression that is left on the community, and so I do not consider dbGAP to be a beacon for data sharing. When we have requested data it has been arduous and left the student or post-doc fatigued and discouraged. I recognize also that the cumbersome process is dictated by agreements and attempts to be backward-compatible with legacy consents. Nonetheless, it has not had nearly the impact it could have had because of these processes, and so it is not a great model for sharing.

3. Policies for the Management and Use of dbGaP Data

- **Alternate controlled-access models**
  I have come to consider dbGAP a "write only" database since data goes in, but rarely comes out. Of course, this is not literally true, but that is the impression that is left on the community, and so I do not consider dbGAP to be a beacon for data sharing. When we have requested data it has been arduous and left the student or post-doc fatigued and discouraged. I recognize also that the cumbersome process is dictated by agreements and attempts to be backward-compatible with legacy consents. Nonetheless, it has not had nearly the impact it could have had because of these processes, and so it is not a great model for sharing.
- **Benefits and risks associated with the availability of genomic study summary statistics**
  The reaction to the Homer et al paper was probably an over-reaction. Summary statistics are probably fine to distribute, and we need to create consents going forward that do not guarantee anonymity to participants, but instead guarantee best effort consistent with scientific utility.
- **NIH is interested in public comments on benefits and risks of such reference use of dbGaP data for research participants, patients, and the scientific community**
  Anything that improves the availability of dbGAP data and its utility should be embraced.

4. General Comments. NIH welcomes general comments on any other topics with regard to dbGaP data submission, access, and management
Sorry to be short and make difficult issues seem easy (they are not) but I wanted to jot down my thoughts quickly before moving to another task. Thank you for putting out this RFI.

Submission Date
02/22/2017
Name
bahram namjou
Primary Purpose of dbGaP Use
Both Study Registration / Data Submission and Data Access / Download
What is your level of experience with dbGaP?
Novice (used dbGaP once, or only a few times)
Role/Other Role
Scientific Researcher
Type of Organization/Other Type
Nonprofit Research Organization
Name of Organization
cchmc

Information Requested

1. dbGaP Study Registration and Data Submission
the process is very smooth and logical. especially the dbgap rep. are very helpful and informative in response to our questions.

2. dbGaP Data Access Request (DAR) and Review
since all downloaded data iare scripted, sometimes it is technically difficult for general researcher to download or open the files after downloading. Some of the files also have very large names that one time, windows (windows 7) was not able to work on the file or re-name it. I have to relocate the file and re-name it in another folder.

3. Policies for the Management and Use of dbGaP Data

- **Alternate controlled-access models**
  none
- **Benefits and risks associated with the availability of genomic study summary statistics**
  after a few years post analyses, there should be a more restrict rules that the person in charge destroy all data.
- **NIH is interested in public comments on benefits and risks of such reference use of dbGaP data for research participants, patients, and the scientific community**
  retrospective gathering information from epic, or medical records is too time consuming for personnel and based on my experience, it doesn't happen in reality unless there is a close collaborations between two parties.

4. General Comments. NIH welcomes general comments on any other topics with regard to dbGaP data submission, access, and management
perhaps, by increasing number of studies, organization of all data will become difficult and need more personnel in NIH.

Submission Date
02/24/2017
Name
Linda Davis
Primary Purpose of dbGaP Use
Study Registration / Data Submission
What is your level of experience with dbGaP?
Never used dbGaP
Role/Other Role
Member of the Public
Type of Organization/Other Type
In Polydactyl study
Name of Organization
Nih Human Genome Study

Information Requested

1. dbGaP Study Registration and Data Submission
Provide info on study results to participants.

2. dbGaP Data Access Request (DAR) and Review
Keep participants in loop on project status.

3. Policies for the Management and Use of dbGaP Data

- **Alternate controlled-access models**
- **Benefits and risks associated with the availability of genomic study summary statistics**
  Sharing data is imperative to success in these studies.
- **NIH is interested in public comments on benefits and risks of such reference use of dbGaP data for research participants, patients, and the scientific community**
  Sharing these results with medical professionals can only help understand and possibly improve care.

4. General Comments. NIH welcomes general comments on any other topics with regard to dbGaP data submission, access, and management.

Submission Date
02/24/2017
Name
Ben Burkley
Primary Purpose of dbGaP Use
Study Registration / Data Submission
What is your level of experience with dbGaP?
Novice (used dbGaP once, or only a few times)
Role/Other Role
Institutional Official
Type of Organization/Other Type
University
Name of Organization
University of Florida

Information Requested

1. dbGaP Study Registration and Data Submission
The single biggest challenge I've experienced is a lack of clear instruction. The dbGaP templates are moderately useful for phenotype submission, however the GWAS data and analysis files are complete chaos. I really have no idea or clear instruction on what they want me to upload. The emails I do receive are very difficult to understand and seem to be off the mark. There should be a single representative/point of contact that is assigned a given study. Ideally this person has very good communication skills and can explain exactly the type of files need to be uploaded.

2. dbGaP Data Access Request (DAR) and Review
The download of data and unpacking of data are an enormous challenge. We had to hire a biotech person in our computing department just to understand all the file extensions, decryption, and unpacking of the data. It was extremely onerous and I've dissuaded other people from using dbGaP.

3. Policies for the Management and Use of dbGaP Data

- **Alternate controlled-access models**
- **Benefits and risks associated with the availability of genomic study summary statistics**
- **NIH is interested in public comments on benefits and risks of such reference use of dbGaP data for research participants, patients, and the scientific community**

4. General Comments. NIH welcomes general comments on any other topics with regard to dbGaP data submission, access, and management

Submission Date
02/25/2017
Name
David Auble
Primary Purpose of dbGaP Use
Study Registration / Data Submission
What is your level of experience with dbGaP?
Novice (used dbGaP once, or only a few times)
Role/Other Role
Scientific Researcher
Type of Organization/Other Type
University
Name of Organization
University of Virginia

Information Requested

1. dbGaP Study Registration and Data Submission
We recently completed the process of registering and submitting data for a study. The entire process took over a year, which is about the length of time that other investigators told me it would take- so our experience is apparently not unusual. dbGaP staff are generally knowledgeable, responsive and helpful, but overall the process is extraordinarily cumbersome. Each document required multiple revisions before being ultimately finalized and accepted. The help documents describe in general what is required, but for someone without prior dbGaP experience, what is required on the various templates and forms is baffling. It would be helpful to have help documents tailored to the different types of studies that are deposited there. For example, our study involved ChIP-seq. The relevant information for such a study differs from that required for a GWAS study. Perhaps dbGaP is not the best home for our type of study, but at the time of submission, it was the best option. Additionally, the interface with our dbGaP site is one-way, meaning that it was difficult or impossible to view and thereby verify what had already been uploaded to the study site. Since the entire process took so long, the inability to see what's there made the process that much more difficult as there was a genuine need to verify materials places on the site months before. Surely there must be a way to expedite the deposition of data in a manner consistent with high ethical and scientific standards such that it takes less than one year to complete. We were lucky that this did not delay publication.

2. dbGaP Data Access Request (DAR) and Review

3. Policies for the Management and Use of dbGaP Data

- **Alternate controlled-access models**
- **Benefits and risks associated with the availability of genomic study summary statistics**
- **NIH is interested in public comments on benefits and risks of such reference use of dbGaP data for research participants, patients, and the scientific community**

4. General Comments. NIH welcomes general comments on any other topics with regard to dbGaP data submission, access, and management

Submission Date
02/27/2017
Name
Elizabeth J Ampleford
Primary Purpose of dbGaP Use
Both Study Registration / Data Submission and Data Access / Download
What is your level of experience with dbGaP?
Novice (used dbGaP once, or only a few times)
Role/Other Role
Scientific Researcher
Type of Organization/Other Type
University
Name of Organization
Wake Forest School of medicine

Information Requested

1. dbGaP Study Registration and Data Submission

2. dbGaP Data Access Request (DAR) and Review
Downloading data: The ability to select the components required would minimize the time and space requirements. In a recent analysis involving SHARP/SHARe I needed the phenotype data and a couple of PLINK files yet I had to wade through masses of raw data to find the files I needed to do the work. In the process I found broken links. The process to inform people of problems is not particularly obvious. I am shuddering at the thought of preparing files for upload.

3. Policies for the Management and Use of dbGaP Data

- **Alternate controlled-access models**
- **Benefits and risks associated with the availability of genomic study summary statistics**
- **NIH is interested in public comments on benefits and risks of such reference use of dbGaP data for research participants, patients, and the scientific community**

4. General Comments. NIH welcomes general comments on any other topics with regard to dbGaP data submission, access, and management

Submission Date
02/27/2017
Name
wade berrettini
Primary Purpose of dbGaP Use
Both Study Registration / Data Submission and Data Access / Download
What is your level of experience with dbGaP?
Experienced (used many times over the course of several years)
Role/Other Role
Scientific Researcher
Type of Organization/Other Type
University
Name of Organization
U. of Pennsylvania

Information Requested

1. dbGaP Study Registration and Data Submission

2. dbGaP Data Access Request (DAR) and Review
My experience with dbGAP until this point has been positive. GWAS chip genotyping and clinical information on 800 DNA samples was submitted to dbGAP by Bill Howells of Washington U. on June 1, 2016. These DNA samples were from a 6 month randomized clinical trial comparing buprenorphine to methadone treatment for opioid addiction. The data were not available to us until February, 2017, a delay of 8 months. This seems excessive. I suspect that dbGAP staffing needs to be increased to provide for smaller delays. The identifier is phs001135.v1.p1

3. Policies for the Management and Use of dbGaP Data

- **Alternate controlled-access models**
- **Benefits and risks associated with the availability of genomic study summary statistics**
- **NIH is interested in public comments on benefits and risks of such reference use of dbGaP data for research participants, patients, and the scientific community**

4. General Comments. NIH welcomes general comments on any other topics with regard to dbGaP data submission, access, and management

Submission Date
02/27/2017
Name
Dongmei Sun
Primary Purpose of dbGaP Use
Both Study Registration / Data Submission and Data Access / Download
What is your level of experience with dbGaP?
Novice (used dbGaP once, or only a few times)
Role/Other Role
Scientific Researcher
Type of Organization/Other Type
University
Name of Organization
University of Alabama at Birmingham Other Type of Organization

Information Requested

1. dbGaP Study Registration and Data Submission
No opinion

2. dbGaP Data Access Request (DAR) and Review
No opinion

3. Policies for the Management and Use of dbGaP Data

- **Alternate controlled-access models**
  No opinion
- **Benefits and risks associated with the availability of genomic study summary statistics**
  No opinion
- **NIH is interested in public comments on benefits and risks of such reference use of dbGaP data for research participants, patients, and the scientific community**
  No opinion

4. General Comments. NIH welcomes general comments on any other topics with regard to dbGaP data submission, access, and management
Since I'm new to dbGAP, I don't have suggestions so far. But I think most people when they first time touch this system, it's a big challenge due to the different background. I hope in the future we can develop some video or documents showing step by step instruction of how to do it. Like information in the link, there are too many info but I didn't figure out a way to know them so far. https://genome.ucsc.edu/cgi-bin/hgTracks?
db=hg19&lastVirtModeType=default&lastVirtModeExtraState=&virtModeType=default&virtMode=0&nonVirtPosition=&position=chr11%3A89224058- 89224058&hgsid=581258651_5mrgknFe0ZOiieZjNNPGvbG7nhvL

Submission Date
02/27/2017
Name
Haixu Tang
Primary Purpose of dbGaP Use Data Access / Download
What is your level of experience with dbGaP?
Novice (used dbGaP once, or only a few times)
Role/Other Role
Scientific Researcher
Type of Organization/Other Type
University
Name of Organization
Indiana University, Bloomington

Information Requested

1. dbGaP Study Registration and Data Submission

2. dbGaP Data Access Request (DAR) and Review
I am not a frequent user of dbGap data; but have been using several dataset in my research. I feel the initial application process is okay; however, the renewal process that is required every year involving the university administration is a bit painful. I am not sure if it can be simplified, at least for some data that are less sensitive.

3. Policies for the Management and Use of dbGaP Data

- **Alternate controlled-access models**
  1) authorized analysis, in which a user submits a computation task (i.e., a program) to a trusted server (on that the sensitive data are stored), and receives the computation results. Only authorized users are allowed to submit programs. The program will be examined for potential malicious code before running, and the computation results will be checked for potential information leak before returning to the user. This model will bring the computation to the data, which will eliminate the risk (and cost) associated with duplicated sensitive data. 2) User query, which can be viewed as a special case of the authorized analysis, where the computation is as simple as one or more queries of specific information in the sensitive data (e.g., a summary statistic). The risks of such queries are relatively easy to control, and thus the user registration can be less stringent.
- **Benefits and risks associated with the availability of genomic study summary statistics**
  The access models mentioned above can provide alternative options for providing access to genomic summary statistics (both can be viewed as the registered access). Note that the registered access also allows for detecting potential malicious users and defending denial-of-service (DoS) attack. The risk mitigation should be prioritized for datasets containing the sensitive information (e.g., HIV infection) and/or vulnerable participants. The methods for mitigating risks with unrestricted access to genomic summary statistics are important research topics, including 1) for monitoring the risks associated with shared data for each participant, and adjusting the sharing strategies accordingly; 2) adding random noises; 3) aggregating multiple datasets to reduce risks; 4) instead of public dissemination, allowing users (either registered or non-registered) to query genomic summary statistics, in which the risks can be dynamically monitored and adjusted.
- **NIH is interested in public comments on benefits and risks of such reference use of dbGaP data for research participants, patients, and the scientific community**

4. General Comments. NIH welcomes general comments on any other topics with regard to dbGaP data submission, access, and management
I hope NIH will continue supporting the research of risk mitigation methods that can be used in sharing genomic summary statistics and other human genomic data.

Submission Date
02/28/2017
Name
Heping Zhang
Primary Purpose of dbGaP Use
Data Access / Download
What is your level of experience with dbGaP?
Experienced (used many times over the course of several years)
Role/Other Role
Scientific Researcher
Type of Organization/Other Type
University
Name of Organization
Yale University

Information Requested

1. dbGaP Study Registration and Data Submission
The data submission process was too complicated and rigid in my own experience. I understand the reasons, but think it adds too much burden on the investigators who submit the data.

2. dbGaP Data Access Request (DAR) and Review
There are so many committees. I can understand the need that a specific committee is needed to review one dataset, but it would be helpful that the approval process is centralized, especially for a continuing use of the several datasets. In other words, one oversight committee may be able to review and approve a renewal request based on the annual report.

3. Policies for the Management and Use of dbGaP Data

- **Alternate controlled-access models**
  I understand the theoretical reasons for various requirements of hosting and downloading the data. I am wondering, however, whether the same protocol has to be applied to all dbGaP datasets. Some of the anonymous databases would be harmless even if they are in the open public domain.
- **Benefits and risks associated with the availability of genomic study summary statistics**
  I am not aware of any additional risk and damage to the study participants as a result from dbGaP data access. The real risks are hyped in my opinion.
- **NIH is interested in public comments on benefits and risks of such reference use of dbGaP data for research participants, patients, and the scientific community**
  As long as the participants' confidentiality is protected, any access should be considered. There can be some cost-sharing for commercial access, which can be used to reward NIH funded investigators for their ideas and efforts.

4. General Comments. NIH welcomes general comments on any other topics with regard to dbGaP data submission, access, and management
dbGaP is a major and great initiative. It would be even better if the process can be more user-friendly.

Submission Date
03/03/2017
Name
Jared Roach
Primary Purpose of dbGaP Use
Both Study Registration / Data Submission and Data Access / Download
What is your level of experience with dbGaP?
Experienced (used many times over the course of several years)
Role/Other Role
Scientific Researcher
Type of Organization/Other Type
Nonprofit Research Organization
Name of Organization
ISB

Information Requested

1. dbGaP Study Registration and Data Submission
There is a massive level of effort for infrequent users of dbGAP to submit data. I tried for over a year to get a dataset submitted, and eventually was not successful. Routine users of dbGAP have automated submission systems, and in some sense dbGAP has evolved to be especially adapted to the data types and formats of routine users. It is increasingly hard for new users or infrequent users to submit data. Because dbGAP submission is tied to funding, dbGAP is therefore a major factor in maintaining funding for groups that already have funding, while denying funding for new groups that may be innovating.

2. dbGaP Data Access Request (DAR) and Review
It is unclear how research not funded by NIH gets into dbGAP. It typically requires a champion inside NIH, and often that is insufficient. And is is unclear how the submission is funded. It us very hard, if not impossible to get data not funded by NIH into dbGAP. Project renewal is a major burden. If it takes an hour per year per dataset, a large analysis requiring multiple datasets may require more than a full day of work every year, just to complete the renewals. The menu listing ALL the dbGAP datasets on a single page during project renewal is confusing. It should just have the relevant datasets. For example, grouped by subject, rather than radio boxes for everything.

3. Policies for the Management and Use of dbGaP Data

- **Alternate controlled-access models**
  Because of the near impossibility of submitting data to dbGAP, and the limited formats and metadata that can be submitted, and the difficulty in getting data from dbGAP, I highly recommend alternate models. For example, ADSP or BSC (bipolar sequencing consortium).
- **Benefits and risks associated with the availability of genomic study summary statistics**
  The benefits of summary statistics such as allele frequencies in particular populations is huge to science and health and has very little downside.
- **NIH is interested in public comments on benefits and risks of such reference use of dbGaP data for research participants, patients, and the scientific community**

4. General Comments. NIH welcomes general comments on any other topics with regard to dbGaP data submission, access, and management

Submission Date
03/04/2017
Name
David Kwiatkowski
Primary Purpose of dbGaP Use
Data Access / Download
What is your level of experience with dbGaP?
Novice (used dbGaP once, or only a few times)
Role/Other Role
Scientific Researcher
Type of Organization/Other Type
Health Care Delivery Organization
Name of Organization
Brigham and Women's Hospital

Information Requested

1. dbGaP Study Registration and Data Submission
Poorly designed process. Needs to be completely revised and redone to simpify the process.

2. dbGaP Data Access Request (DAR) and Review
Terribly designed and implemented process. The requirement for review by institutional officials is a huge waste of time, since my IOs are busy with other tasks. Takes hours/days to access one study. Simplify the process!

3. Policies for the Management and Use of dbGaP Data

- **Alternate controlled-access models**
  Yes, simplify the process of access! Currently it is a waste of time, and impedes scientific progress.
- **Benefits and risks associated with the availability of genomic study summary statistics**
  Simplify and facilitate access!
- **NIH is interested in public comments on benefits and risks of such reference use of dbGaP data for research participants, patients, and the scientific community**
  Yes, we'd all like simpler access.

4. General Comments. NIH welcomes general comments on any other topics with regard to dbGaP data submission, access, and management

Submission Date
03/10/2017
Name
Ken Winters
Primary Purpose of dbGaP Use
Browsing Unrestricted-Access Study Information
What is your level of experience with dbGaP?
Novice (used dbGaP once, or only a few times)
Role/Other Role
Scientific Researcher
Type of Organization/Other Type
Nonprofit Research Organization
Name of Organization
Oregon Research Institute

Information Requested

1. dbGaP Study Registration and Data Submission
The dbGaP study registration process was generally clear. And the dbGaP staff were responsive and helpful with questions.

2. dbGaP Data Access Request (DAR) and Review

3. Policies for the Management and Use of dbGaP Data

- **Alternate controlled-access models**
- **Benefits and risks associated with the availability of genomic study summary statistics**
- **NIH is interested in public comments on benefits and risks of such reference use of dbGaP data for research participants, patients, and the scientific community**
  The potential for dbGaP to support research is significant, particularly given the need for very large data sets with candidate gene and linkage studies.

4. General Comments. NIH welcomes general comments on any other topics with regard to dbGaP data submission, access, and management
Perhaps consider adding a section to the Data Submission form to indicate which PhenX assessment tools were used (if any) in measuring phenotypes.

Submission Date
03/17/2017
Name
Jessie Tenenbaum
Primary Purpose of dbGaP Use
Data Access / Download
What is your level of experience with dbGaP?
Novice (used dbGaP once, or only a few times)
Role/Other Role
Scientific Researcher
Type of Organization/Other Type
University
Name of Organization
Duke University

Information Requested

1. dbGaP Study Registration and Data Submission

2. dbGaP Data Access Request (DAR) and Review
I only used dbGap intensively once, and it's been several months now so some of the details have faded, but a few comments-
1. at a few different points, there was a lot of red tape to deal with that project members OTHER than the PI needed to handle BUT dbGap only allowed it to be done in the PI's account. PIs should be able to designate personnel to handle more steps than they currently do. (unfortunately, i don't remember the specific tasks from the DAR part. Project renewal is one of these areas.)
2. Downloading data- we were looking at Framingham, which is HUGE. It was difficult first to figure out which subsets we needed to request. Once we did that, it was surprising and a little confusing that we then had access to ALL Framingham data. I think/hope we didn't download anything we were't supposed to, but I'm not 100% sure.

3. Policies for the Management and Use of dbGaP Data

- **Alternate controlled-access models**
  One huge obstacle we had (have) is that we're a consortium of many institutions doing analysis on the same dataset. It's a huge pain to have to have EACH institution get its own IRB approval/exemption, fill out a DAR, get the signing officer involved. Some rules for responsible sharing of data with collaborators at other institutions would have been much easier. (we still don't yet have access for anyone beyond Duke.)
- **Benefits and risks associated with the availability of genomic study summary statistics**
- **NIH is interested in public comments on benefits and risks of such reference use of dbGaP data for research participants, patients, and the scientific community**
  I am all in favor, but may be biased as a researcher (and a generally very open person willing to share a lot of personal information).

4. General Comments. NIH welcomes general comments on any other topics with regard to dbGaP data submission, access, and management

Submission Date
03/21/2017
Name
Cathy C. Laurie
Primary Purpose of dbGaP Use
Both Study Registration / Data Submission and Data Access / Download
What is your level of experience with dbGaP?
Experienced (used many times over the course of several years)
Role/Other Role
Scientific Researcher
Type of Organization/Other Type
University
Name of Organization
University of Washington

Information Requested

1. dbGaP Study Registration and Data Submission
Registration: This is a lengthy process involving local IRB review of original consent documents. However, it is not clear how to shorten it while also maintaining high standards of respect for participant consent and Data Use Limitations. Data submission: a. Submission of data files through the submission portal is convenient and straightforward. The option for uploading via Aspera command line (“ascp”) is very useful. b. In studies that have parent and child accessions, it would be helpful to have documentation that clarifies what file types go into the parent versus the child accession(s). c. As dbGaP continues to grow and participants get included in multiple genotyping and sequencing efforts, the degree of overlap between studies will also grow. It would be useful for dbGaP to clarify what responsibility submitters have for finding out what other submitted datasets samples belong to and annotating these overlaps in the Subject Consent file. Standardization of data formats: When working with downloaded text files (e.g., sample-subject mapping and phenotype files), more standardization of file formats would facilitate automated processing on the user end. For example, some (but not all) files have extra delimiters at the end of each line that do not have a corresponding header. In other cases, the headers on the data in required files have names that are not standardized. For example, “shareid”, “subjid”, and “SUBJECT_ID” may be the subject ID column in the Subject files from different studies; ideally, they would all have “SUBJECT_ID” as the header for column corresponding to subject ID. These issues seem to affect mainly older accessions and it would helpful to have some additional curation of these older studies as resources allow.

2. dbGaP Data Access Request (DAR) and Review
Data Access Requests: a. When beginning a new project, it is a bit confusing to select a single accession as a first step, followed by the general project descriptions, but then coming back to select additional accessions. b. When submitting a DAR for a new accession to be added to an existing project, it appears that all previously selected (and currently approved) accessions also get resubmitted. When the project contains many different accessions, it is very time-consuming to go through and re-check all of the DULs, etc. Can the paperwork be limited to new requests? c. In my experience, IRB renewals are required 3-4 months before they expire, which means that they must be renewed every 8-9 months rather than yearly (for IRB approvals on a yearly schedule). Is this policy necessary? In any case, can submitters get a warning that their IRB needs to be renewed prior to submission (rather than getting the message later from the DAC)? d. Regarding data set selection: it would be useful to be able to select an entire set of data sets that belong to a consortium (e.g. TOPMed) simultaneously, rather than one at a time. Identifying data sets of interest: The advanced search capabilities in dbGaP are very good. The facets search interface is particularly helpful and intuitive. Making the facets search more prominent than the Entrez search would be helpful. Updated tutorials for how to use the different search functions, and a clearer distinction between the Entrez and facets searches would be very useful. Data Access Review Process a. The DAR review process is generally done quickly, but if anything is missing from the rather complicated application, it can cause a long delay. Perhaps automated checks prior to submission could help this situation. b. DAC review is an important component of respecting participant consent and it seems to be effective in that regard. Downloading data: When downloading data from the Access Request page, displaying the entire folder hierarchy (rather than clicking on the plus buttons to expand each folder) could reduce the time spent waiting for the file listing to load.

3. Policies for the Management and Use of dbGaP Data

- **Alternate controlled-access models**
  The dbGaP serves data in units corresponding to study-consent groups, which is directly related to how data are requested and reviewed for compliance with participant consent. Distribution of data in separate chunks by study-consent group entails extra work for investigators who wish to run combined analyses across such units, as they need to identify appropriate files and piece them together. An alternative approach to data distribution might be to serve the data within a cloud environment where a user might be given access to data sets that include multiple study-consent group units. The system could, in principle, serve the data from a relational database that would allow users to search for what they need and combine data according to their permissions (i.e. approved DARs). This would, of course, entail substantial new infrastructure development by dbGaP, requiring new funding.
- **Benefits and risks associated with the availability of genomic study summary statistics**
- **NIH is interested in public comments on benefits and risks of such reference use of dbGaP data for research participants, patients, and the scientific community**

4. General Comments. NIH welcomes general comments on any other topics with regard to dbGaP data submission, access, and management

The dbGaP and associated NIH Data Access Committees have performed an invaluable service to the scientific community by curating, storing and distributing large datasets via controlled access. The system has been heavily used, but to my knowledge, there have been no significant incidents in which participant consent was seriously violated in the use of data obtained through dbGaP. Therefore I think that the system has been very effective in ensuring compliance with participant consent. Nevertheless, there are time-consuming and cumbersome aspects to the system, many of which could be streamlined with sufficient support from NIH for these functions. As a Data Coordinating Center, my group has worked extensively with study investigators and dbGaP to support data submission and curation on more than 50 projects. The dbGaP team members and their leader (Mike Feolo) do a great job in this area with limited resources. They are extremely helpful and accommodating to special needs of studies, especially large consortia like TOPMed. They also do an excellent job in curating data submitted by studies. This often requires extensive communication back and forth with investigators, and multiple rounds of data submission, to obtain complete data in the requested format. My impression is that the requirements of this effort are under-appreciated.

Submitting study investigators also do a tremendous service to the community and are generally not supported sufficiently for the work that is required to provide clean, well-documented data in the required format. Therefore, additional NIH support is needed for both study investigators and dbGaP to make high quality data available quickly to the general scientific community.

Submission Date
03/22/2017
Name
Richard Shaw (& Matthew Young)
Primary Purpose of dbGaP Use Data Access / Download
What is your level of experience with dbGaP?
Novice (used dbGaP once, or only a few times)
Role/Other Role
Scientific Researcher
Type of Organization/Other Type
Biotech/Pharmaceutical Company
Name of Organization
Repositive Ltd.

Information Requested

1. dbGaP Study Registration and Data Submission
N/A

2. dbGaP Data Access Request (DAR) and Review
In 2015, my former colleague Matthew Young applied for General Research Use consent group access to dbGaP to test methods developed for efficient storage and querying of human genomic variants from thousands of samples with privacy preservation of participants. This was Repositive's initial application for access to dbGaP data. It took him a couple of months and he summarised his experience of the entire application process from scratch in the following blog posts :-
https://blog.repositive.io/watching-paint-dry-in-the-21st-century-or-applying-for-data-from-dbgap/
https://blog.repositive.io/accessing-dbgap-a-bureaucratic-oddesey-part-2/
https://blog.repositive.io/how-to-successfully-apply-for-access-to-dbgap/ As Matthew's replacement on the project (following his departure from the company), I had assumed that a PI name change by the Signing Official would suffice for me to take over his dbGaP access responsibilities. Instead I found that I had to arrange for him to close out his access and create a new access request for the same project, thus repeating a month of his experience.

3. Policies for the Management and Use of dbGaP Data

- **Alternate controlled-access models**
  N/A
- **Benefits and risks associated with the availability of genomic study summary statistics**
  N/A
- **NIH is interested in public comments on benefits and risks of such reference use of dbGaP data for research participants, patients, and the scientific community**
  N/A

4. General Comments. NIH welcomes general comments on any other topics with regard to dbGaP data submission, access, and management
We appreciate the service that dbGaP provides and the work of the DACs and hope that this feedback will assist in improving the experience of other dbGaP users.

Submission Date
03/28/2017
Name
N/A
Primary Purpose of dbGaP Use
Study Registration / Data Submission
What is your level of experience with dbGaP?
Experienced (used many times over the course of several years)
Role/Other Role
Bioinformatics Programmer
Type of Organization/Other Type
University
Name of Organization
Baylor College of Medicine

Information Requested

1. dbGaP Study Registration and Data Submission
The one comment that I have relates the the sample registration files. Sometimes, another collaborator will submit the sample registration files without my input. I will then submit all of the data files (usually BAMs and VCFs). If there is then issues with the sample registration files, I am not able to look at them to try and resolve the issue. It would be very helpful as a submitter to have access to these files on the web portal. Currently, I can replace the existing files with new ones, but cannot view the files themselves.

2. dbGaP Data Access Request (DAR) and Review

3. Policies for the Management and Use of dbGaP Data

- **Alternate controlled-access models**
- **Benefits and risks associated with the availability of genomic study summary statistics**
- **NIH is interested in public comments on benefits and risks of such reference use of dbGaP data for research participants, patients, and the scientific community**

4. General Comments. NIH welcomes general comments on any other topics with regard to dbGaP data submission, access, and management

Submission Date
04/02/2017
Name
Edwin K. Silverman
Primary Purpose of dbGaP Use
Both Study Registration / Data Submission and Data Access / Download
What is your level of experience with dbGaP?
Experienced (used many times over the course of several years)
Role/Other Role
Scientific Researcher
Type of Organization/Other Type
Health Care Delivery Organization
Name of Organization
Brigham and Women's Hospital

Information Requested

1. dbGaP Study Registration and Data Submission

2. dbGaP Data Access Request (DAR) and Review
I appreciate the opportunity to provide comments regarding dbGaP data submission, access, and management procedures. My research group relies on dbGaP for many aspects of our scientific investigations, including providing a safe, accessible, monitored access location for phenotype and genotype information generated in studies in which we participate, as well as a valuable resource to obtain phenotype and genotype data from other studies for replication efforts. I have always found the dbGaP staff to be extremely responsive and helpful. However, despite their efforts, I believe that the Data Access Request process remains a major hindrance to data sharing. It is just too long and too complex. A couple of key issues: 1) Some studies still require IRB approval to get access to their dbGaP data. This typically requires an application to the IRB, which is followed by the IRB deciding that this research with deidentified data does not qualify as human subjects research, and finally the necessary IRB waiver/approval is obtained. However, this cycle can take months to complete. It would be helpful if those studies that require IRB approval could revisit this issue with their IRBs to determine if this additional approval is still required—and why. 2) At our institution, obtaining sign-off from the IT Director can take a long time. If there could be institutional-level IT approval that was updated on an annual basis—rather than requiring each separate data request to have IT approval—it would substantially streamline the data request process. 3) Another hurdle to obtaining data through dbGaP is the requirement for sign-off by an institutional official. If there is any way that responsibility for appropriate use of the data could rely with the principal investigator without institutional sign-off, I suspect that the efficiency of the data access request process would improve substantially. 4) There is wide variability in the approaches used to submit phenotype data from different studies. In many cases, the deposited phenotypic data is confusing and difficult for someone outside of the parent study to use. A more standardized, simplified approach for phenotypic data organization would assist other investigators in the proper use of the deposited data sets.

3. Policies for the Management and Use of dbGaP Data

- **Alternate controlled-access models**
- **Benefits and risks associated with the availability of genomic study summary statistics**
- **NIH is interested in public comments on benefits and risks of such reference use of dbGaP data for research participants, patients, and the scientific community**

4. General Comments. NIH welcomes general comments on any other topics with regard to dbGaP data submission, access, and management

Submission Date
04/03/2017
Name
Ara Tahmassian
Primary Purpose of dbGaP Use
What is your level of experience with dbGaP?
Role/Other Role
Institutional Official
Type of Organization/Other Type
University
Name of Organization
Harvard University

Information Requested

1. dbGaP Study Registration and Data Submission

2. dbGaP Data Access Request (DAR) and Review

3. Policies for the Management and Use of dbGaP Data

- **Alternate controlled-access models**
- **Benefits and risks associated with the availability of genomic study summary statistics**
- **NIH is interested in public comments on benefits and risks of such reference use of dbGaP data for research participants, patients, and the scientific community**

4. General Comments. NIH welcomes general comments on any other topics with regard to dbGaP data submission, access, and management

Additional Document

Office of Science Policy (OSP)
National Institutes of Health
6705 Rockledge Drive, Suite 750
Bethesda, MD 20892

Re: Request for Information on Processes for dbGaP Data Submission, Access, and Management

Notice Number: NOT-OD-17-044

This response is being submitted on behalf of Harvard University. The commenters include faculty/researchers, the authorized institutional signing officials responsible for reviewing and approving the Data Use Certifications, Institutional Officials, Institutional Review Board members, and information technology and security officers.

1. **dbGaP Study Registration and Data Submission**

   a) Over all the data submission and registration process appears to be working well. One concern is that there may be additional criteria for evaluating voluntary deposits of data that are not known in advance and because voluntary deposits of data are given a lower priority for review than required deposits of data, the lag time between the reviews of the two types of deposits can be problematic. For instance, it is not yet clear whether the registry will accept data related to stem cell lines that are not on the NIH registry. Because of the potential that voluntary data submissions could be denied, researchers who plan to share valuable data with colleagues by making it publicly available, or who are required to deposit the data for publication purposes, must seek alternative data-repositories to make the data from their voluntary submission available by the time of publication. The use of alternative data-repositories fragments the distribution of the locations of available data, and limits the aggregation of data for those seeking information. It also has the additional burden of identifying multiple data-sources. *We strongly encourage NIH to consider voluntary deposits of data at the same time as required deposits of data when all of the data is related to the same publication and it would be valuable if there were more guidance in regards to when voluntary deposits may not be accepted*.

   b) There are some distinctions in the way an institution and/or an IRB should review different types of large scale human genomic studies. For example, a GWAS study where the provenance documentation is certified in advance of the collection is distinct from a study using previously established hESC lines from a number of different institutions and collection protocols; in the latter it has been suggested that cell line names should be included in the Institutional Certifications, which makes sense but was not apparent through published guidance. Additionally, the “Points to Consider Page” linked to from NIH Guidance for Non-NIH funded studies exclusively addresses GWAS. *Making additional guidance on distinct requirements for different types of studies would be welcomed.*

   c) The initial step in the dbGaP submission process is for the PI to email the NIH Program Officer or the appropriate Genomic Program Administrator with the requisite information and documentation for the data being submitted. It then resides with that one individual and only once the Officer or Administrator registers the study is the PI invited to the system to enter the study and complete the registration and submission. *It would be valuable if there were an application portal that allows researchers and their staff to input the initial request instead of emailing it because that would streamline communications and create a system of record. A single system of record at this stage would allow researchers to have their staff prepare the application for them without many emails back and forth and it would also give researchers immediate confirmation that their applications were submitted so they do not have to wait for email confirmation. Also, this would reduce redundancies in having to upload some of the data later in the registration process. We also suggest that at the time of submitting the application it would be helpful to let researchers know the estimated turnaround time for the review.*

2. dbGaP Data Access Request (DAR) and Review

a. DAR Process – credentialing and routing

The system currently leverages eRA Commons log-in credentials, which is helpful since the majority, if not all, of our faculty already have these credentials. However, there are others such as post-doctoral fellows and students who have the PI role associated with their Commons ID as a result of fellowship awards/submissions but would not have the appropriate PI status needed in order to submit a DAR. Currently, there is no system limitation put in place to prevent these individuals from submitting a request. These are then only noticed when the DAR is routed to the institutional official for signature, and the DAR must be pulled back and resubmitted under the faculty mentor. *An appropriate system prompt or limitation could help alleviate the wasted time and effort of these requests*.

The routing of DARs to the institutional official follows rules similar to the Business Official routing used in the xTrain system. Like xTrain, only the institutional official to whom the request is routed can approve/submit the request. Unlike xTrain, however, only the institutional official is able to even view the request. Therefore, the Institutional Official and not a designee is required to verify the veracity and appropriateness of the request, a process that could be delegated appropriately. Further, only one person in the institution has the access to the details about a given request, which can create significant difficulties when/if that person is unavailable or leaves the institution. If the DAR is routed to that individual, currently the process requires that requester recall the request and re-submit to the substitute contact, a process that further delays submission of the request and entails additional administrative burden. *Routing similar to that used for the Research Performance Progress Report (RPPR) would be far preferable. This would allow the requester to select the appropriate individual to receive an email notification that the DAR had been submitted but anyone at the institution with the appropriate Commons role would be able to view and submit the request*.

b. DAR Process – the Data Use Certification(s)

Currently, each dbGaP data set requested as part of a project contributes its own unique 6-7 page Data Use Certification (DUC) to the package that must be institutionally reviewed and approved. PIs are able to request access to multiple data sets for a single project in a single request, which helps reduce the administrative burden on the researcher. However, the current DUC structure presents an unusually high burden to the institutional official. Additionally, similar to the concerns that led NIH to move away from publication embargoes in the new GDS policy, this also makes it difficult, if not impossible, for PIs to keep track of and comply with different terms that may apply to the multiple data sets they are using for a single project.

There are also some data sets, for example the Alzheimer's Disease Genetics Consortium data that require supplemental agreements to be submitted in addition to the DUC. To the extent possible, any special terms for a given data set should be limited to those absolutely required by law, regulation, or the original consent forms, and if any such special terms are necessary, they should be incorporated into the system generated DUC. The supplemental agreements add significant additional burden by requiring a signing official to review and sign the supplemental agreement and then separately review and sign the DUC. Additionally, we have encountered examples where a request was rejected because the supplemental agreement had been updated by the investigators only; this leads to additional delays and burden in reviewing and signing the new supplemental agreement so that the request could be re-submitted.

Large sections of the text are either very similar or exactly the same in each DUC. There are some key sections that do appear to differ, such as the text to be utilized in acknowledging use of the data set or references to the Data Access Committee applicable for the given data set. *Revising the DUC so that there is one set of common core terms for all data sets with a section that clearly outlines any unique requirements for a specific data set would significantly reduce administrative burden while facilitating compliance*.

The terms incorporated into the DUC should, to the extent possible, be consistent with the terms included in the data use agreements used by other NIH repositories providing access to similarly controlled data. Currently, each NIH repository has a different agreement, which places undue burden in the review and signature of these agreements and in compliance. *The greater consistency across definition and terms, the easier it will be for institutions to comply*.

c. Data Access Committee Review

For some requests, we have experienced very lengthy delays (up to a couple months or more) waiting for the results of the review. The worst-case scenario is when the request is ultimately rejected for an administrative item that could have easily been addressed earlier, such as omission in uploading the updated IRB approval letter for a renewal. *Enhanced compliance checks could be incorporated into the system to ensure that all administrative components are in place prior to DAC review so that the DAC can focus on review for compliance with any Use Limitation. Additionally, where the Use Limitation allows for very broad use of the data, perhaps an "expedited" review process could be utilized*. Further, it would be helpful to have expectations for turnaround times for review.

d. Downloading Data from dbGaP

For very large data sets, such as The Cancer Genome Atlas (TCGA), the length of time required to download the data from dbGaP (e.g. several weeks or more) can be prohibitive to investigators wishing to use the data. Additionally, the data set then consumes a large amount of storage space at the institution. *We propose that NIH consider that the data sets be accessed via VPN and/or from the cloud, allowing investigators more rapid access to the data and leading to a corresponding increase in use of the data. Alternatively, dbGaP could allow institutions to retain copies of frequently used data beyond the active approved project. The institution could certify that no research access to the data set would be allowed without an active, approved project, enabling the next researcher to have immediate access to the data set upon approval by dbGaP*.

e. Project renewal and close-out processes

Currently, it is not possible to determine from the project record in the dbGaP system when the annual renewal or close-out of the project is due. The record provides the expiration date for approval of each data set included in the project, but we often receive reminders that renewal or close-out are due for projects by a date that is not reflective of any of those dates. *We recommend that a single expiration date for the project to be listed in the record, which would make it easier for the institution to manage compliance with this requirement*.

The requirement for annual renewal reports is also unnecessarily burdensome. Projects utilizing the dbGaP data are often also funded by a grant from the NIH, which generally have a period of performance of 2-5 years. *The renewal period should, correspondingly, be set with the expiration of the grant*. As part of the application process, the Investigator could be given the option to select a requested initial period, up to an appropriate maximum amount, after which renewal would be required.

The contributors include: Caroline Shamu (NIH LINCS Program U54 HL127365, U54HL127624-03S2); Chris Botka, Director of Research Computing, HMS; Ara Tahmassian, University Chief Research Compliance Officer and IO; Kirsten Elwell, Department of Stem Cell and Regenerative Biology; Melissa J. Lopes, Senior Research Compliance Officer; Melissa Korf (Associate Director, Grants & Contracts); Leslie Moser Howes, Managing Director, Office of Human Research Administration-Longwood Area; Shannon Sewards, Director of IRB Administration-University Area; and Barbara E. Bierer, MD, Director, Regulatory Foundations, Ethics and Law Program, Harvard Catalyst (Harvard CTSA).

Respectfully submitted on behalf of Harvard University

Submission Date
04/03/2017
Name
Lasse Folkersen
Primary Purpose of dbGaP Use
Both Study Registration / Data Submission and Data Access / Download
What is your level of experience with dbGaP?
Novice (used dbGaP once, or only a few times)
Role/Other Role
Scientific Researcher
Type of Organization/Other Type
University
Name of Organization
Technical University of Denmark

Information Requested

1. dbGaP Study Registration and Data Submission

2. dbGaP Data Access Request (DAR) and Review
The data Access to dbGap is ridiculously complicated, as I'm sure you are already aware of. I suggest you streamline it, and one easy place to start is to get rid of all the managing/administrative level OKs. Just hold individual researchers individually responsible.

3. Policies for the Management and Use of dbGaP Data

- **Alternate controlled-access models**
- **Benefits and risks associated with the availability of genomic study summary statistics**
  I think you are being overly cautious with per-snp summary level data. Those papers that have shown a risk of privacy breach are too contrived to be a real problem.
- **NIH is interested in public comments on benefits and risks of such reference use of dbGaP data for research participants, patients, and the scientific community**

4. General Comments. NIH welcomes general comments on any other topics with regard to dbGaP data submission, access, and management
Whatever you do, do it fast because the field is already fragmenting as researchers require and develop less bureaucratic solutions.

Submission Date
04/05/2017
Name
Phil Canakis
Primary Purpose of dbGaP Use
Both Study Registration / Data Submission and Data Access / Download
What is your level of experience with dbGaP?
Never used dbGaP
Role/Other Role
Manager IT Service and Support, Research Informatics
Type of Organization/Other Type
University
Name of Organization
The Pennsylvania State Univ Hershey Med Ctr

Information Requested

1. dbGaP Study Registration and Data Submission
Feedback from our researchers: Staff scientists and experienced PhD scholars should be permitted to initiate studies regardless of academic title.

2. dbGaP Data Access Request (DAR) and Review
Feedback from our researchers: There should be a mechanism for pre-approval for multiple years instead of requiring yearly renewals. When it is known that a project will span multiple years, the additional repetitive steps for approval are unnecessary and time consuming.

3. Policies for the Management and Use of dbGaP Data

- **Alternate controlled-access models**
  Feedback from our researchers: • Allow download of subset of targeted coordinates representing genes of interest instead of requiring full data set downloads. The lack of genes/intervals query capability forced researchers and IT to spend unnecessary time and resources to complete the project goals. • While identified data is appropriately restricted and controlled, most of the restrictions for de-identified should be relaxed to ease requesting data sets, data usage, downloading protocols, and storage protocols. • Additional documentation is needed on data download and processing. User support through the NIH helpdesk was confusing and time consuming. • Documentation and download mechanisms are geared towards users that have an in-depth information and technology background.
  • The portal did not make it easy to determine the file size and download time and lacked a proper search function.
  • Meta descriptors are hard to use.

- **Benefits and risks associated with the availability of genomic study summary statistics**
- **NIH is interested in public comments on benefits and risks of such reference use of dbGaP data for research participants, patients, and the scientific community**

4. General Comments. NIH welcomes general comments on any other topics with regard to dbGaP data submission, access, and management
Feedback from our researchers: Formal or informal tutorial system or training would be useful. We solicited information from our investigators to respond collectively to this RFI.

Submission Date
04/05/2017
Name
FASEB
Primary Purpose of dbGaP Use
What is your level of experience with dbGaP?
Role/Other Role
Professional Association
Type of Organization/Other Type
Professional Org/Association
Name of Organization
Federation of American Societies for Experimental Biology

Information Requested

1. dbGaP Study Registration and Data Submission

2. dbGaP Data Access Request (DAR) and Review

3. Policies for the Management and Use of dbGaP Data

- **Alternate controlled-access models**
  FASEB supports continued use of controlled access for whole genome sequences and any other datasets or data products that pose a risk for re- identification of research participants. When selecting a model for controlled-access, we encourage NIH to consider approaches that minimize administrative burden for researchers depositing and accessing data. Similarly, NIH also should ensure that the adopted model is uniformly implemented across all NIH genomic databases.
- **Benefits and risks associated with the availability of genomic study summary statistics**
  FASEB recognizes that the exchange of research findings increases the efficiency of scientific research and can accelerate discoveries that improve human health. We appreciate NIH's consideration of alternative access models for data products that present a much lower risk for re-identification than individual-level data. Below we describe two types of risk and suggest mitigating them through use of registered- and controlled-access models. First, accurate interpretation of genomic data and associations are challenging due to their complexity and many potential complicating factors. FASEB recommends a registration model that incorporates a brief educational module for first-time users. The educational component could be tailored to the type of user, with different versions for clinicians, researchers, and members of the public. For researchers, this training should emphasize rigorous and reproducible research practices for using these data products. The modules should also affirm the ethical and legal responsibilities of users to maintain the privacy of research participants and clearly state that users are prohibited from attempting to re-identify human subjects. Registration and training should be transferable across all NIH genomic databases. Second, providing greater access to summary statistics increases risks for some research participants. In limited circumstances, summary statistics could be used to determine whether a particular individual participated in a particular study; the nature of the study could, in turn, reveal further health information about that individual. There is also concern that summary statistics from research on sensitive phenotypes and vulnerable populations could be misused to further marginalize these groups. Unless these risks were explicitly covered in the informed consent process, summary statistics should only be available through a controlled-access system for any studies involving sensitive phenotypes, vulnerable populations, children, or smaller sample sizes.
- **NIH is interested in public comments on benefits and risks of such reference use of dbGaP data for research participants, patients, and the scientific community**
  Allele frequencies and other summary statistics can help inform the interpretation of clinical test results. However, FASEB is concerned about the potential harm from misinterpretation and misapplication of such summary information – especially in the context of healthcare decisions. Information that could mislead clinicians and patients may inadvertently reduce patient safety. As described in the prior section, we advocate that NIH implement a short educational module as part of the registration process for accessing dbGaP summary statistics. The module should describe best practices for clinical use and introduce users to concepts such as correlation versus causation, effect size, probability of significance, and the risks of generalizing results to other populations.

4. **General Comments. NIH welcomes general comments on any other topics with regard to dbGaP data submission, access, and management**

Additional Document

FASEB
Federation of American Societies
for Experimental Biology

Representing 125,000 Researchers

301.634.7000
www.faseb.org

9650 Rockville Pike
Bethesda, MD 20814

April 5, 2017

Office of Science Policy
National Institutes of Health
6705 Rockledge Drive, Suite 750
Bethesda, MD 20892

Dear NIH Science Policy Team:

The Federation of American Societies for Experimental Biology (FASEB) appreciates the opportunity to provide comments in response to the National Institutes of Health's (NIH's) "Request for Information on Processes for dbGaP Data Submission, Access, and Management" (NOT-OD-17-044). FASEB is composed of 30 scientific societies representing 125,000 biomedical and biological investigators. We commend NIH for seeking stakeholder input to enhance the Database of Genotypes and Phenotypes (dbGaP) and offer cautious support for expanding access to summary statistics. To balance potential risks and benefits, we recommend use of a registration model with an educational component for accessing most types of summary statistics. However, for studies involving sensitive phenotypes, vulnerable populations, children, or smaller sample sizes, we recommend that summary statistics only be available through a controlled-access system. Below we provide specific details in responses to three issues raised in the RFI.

Alternate controlled-access models

FASEB supports continued use of controlled access for whole genome sequences and any other datasets or data products that pose a risk for re-identification of research participants. When selecting a model for controlled-access, we encourage NIH to consider approaches that minimize administrative burden for researchers depositing and accessing data. Similarly, NIH also should ensure that the adopted model is uniformly implemented across all NIH genomic databases.

Benefits and risks associated with the availability of genomic study summary statistics

FASEB recognizes that the exchange of research findings increases the efficiency of scientific research and can accelerate discoveries that improve human health. We appreciate NIH's consideration of alternative access models for data products that present a much lower risk for re-identification than individual-level data. Below we describe two types of risk and suggest mitigating them through use of registered- and controlled-access models.

First, accurate interpretation of genomic data and associations are challenging due to their complexity and many potential complicating factors. <u>FASEB recommends a registration model that incorporates a brief educational module for first-time users</u>. The educational component could be tailored to the type of user,

The American Physiological Society • American Society for Biochemistry and Molecular Biology • American Society for Pharmacology and Experimental Therapeutics
American Society for Investigative Pathology • American Society for Nutrition • The American Association of Immunologists • American Association of Anatomists
The Protein Society • Society for Developmental Biology • American Peptide Society • Association of Biomolecular Resource Facilities
The American Society for Bone and Mineral Research • American Society for Clinical Investigation • Society for the Study of Reproduction • The Teratology Society
The Endocrine Society • The American Society of Human Genetics • International Society for Computational Biology • American College of Sports Medicine
Biomedical Engineering Society • Genetics Society of America • American Federation for Medical Research • The Histochemical Society • Society for Pediatric Research
Society for Glycobiology • Association for Molecular Pathology • Society for Redox Biology and Medicine • Society For Experimental Biology and Medicine
American Aging Association (AGE) • U. S. Human Proteome Organization (US HUPO)
32

with different versions for clinicians, researchers, and members of the public. For researchers, this training should emphasize rigorous and reproducible research practices for using these data products. The modules should also affirm the ethical and legal responsibilities of users to maintain the privacy of research participants and clearly state that users are prohibited from attempting to re-identify human subjects. Registration and training should be transferable across all NIH genomic databases.

Second, providing greater access to summary statistics increases risks for some research participants. In limited circumstances, summary statistics could be used to determine whether a particular individual participated in a particular study; the nature of the study could, in turn, reveal further health information about that individual. There is also concern that summary statistics from research on sensitive phenotypes and vulnerable populations could be misused to further marginalize these groups. Unless these risks were explicitly covered in the informed consent process, summary statistics should only be available through a controlled-access system for any studies involving sensitive phenotypes, vulnerable populations, children, or smaller sample sizes.

Clinical Use of Genomic Research Data Maintained in Controlled-Access in dbGaP

Allele frequencies and other summary statistics can help inform the interpretation of clinical test results. However, FASEB is concerned about the potential harm from misinterpretation and misapplication of such summary information – especially in the context of healthcare decisions. Information that could mislead clinicians and patients may inadvertently reduce patient safety. As described in the prior section, we advocate that NIH implement a short educational module as part of the registration process for accessing dbGaP summary statistics. The module should describe best practices for clinical use and introduce users to concepts such as correlation versus causation, effect size, probability of significance, and the risks of generalizing results to other populations.

FASEB appreciates NIH’s engagement of the biomedical community as it explores ways to enhance dbGaP. We are supportive of expanded access to summary statistics provided that (1) there is an educational component on appropriate and rigorous use; and (2) summary statistics from more sensitive studies remain only available through a controlled-access system. Please do not hesitate to contact me if FASEB can provide further assistance.

Sincerely,

Hudson Freeze

Hudson H. Freeze, PhD
FASEB President

Submission Date
04/06/2017
Name
Preston Campbell, CEO
Primary Purpose of dbGaP Use
Both Study Registration / Data Submission and Data Access / Download
What is your level of experience with dbGaP?
Experienced (used many times over the course of several years)
Role/Other Role
Scientific Researcher
Type of Organization/Other Type
Non-Profit Research Funding Organization
Name of Organization
Cystic Fibrosis Foundation

Information Requested

1. dbGaP Study Registration and Data Submission
Please see attached comment document.

2. dbGaP Data Access Request (DAR) and Review
Please see attached comment document.

3. Policies for the Management and Use of dbGaP Data

- **Alternate controlled-access models**
  Please see attached comment document.
- **Benefits and risks associated with the availability of genomic study summary statistics**
  Please see attached comment document.
- **NIH is interested in public comments on benefits and risks of such reference use of dbGaP data for research participants, patients, and the scientific community**
  Please see attached comment document.

4. General Comments. NIH welcomes general comments on any other topics with regard to dbGaP data submission, access, and management
Please see attached comment document.

Additional Document

CYSTIC FIBROSIS
FOUNDATION
ADDING TOMORROWS

April 6, 2017

Carrie D. Wolinetz, Ph.D.
Director
Office of Science Policy
National Institutes of Health
6705 Rockledge Drive
Bethesda, MD 20817

Re: Notice No. NOT-OD-17-044, Request for Information on Processes for dbGaP Data Submission, Access, and Management

Dear Director Wolinetz:

The Cystic Fibrosis Foundation appreciates the opportunity to provide comments in response to the Office of Science Policy's request for information on the data submission, management, and access processes for the NIH National Center for Biotechnology Information (NCBI) database of Genotypes and Phenotypes (dbGaP). dbGaP is a critical resource for researchers and sets a framework of key standards and much needed infrastructure for data sharing in the research community.

The Cystic Fibrosis Foundation strongly supports facilitating and supporting data sharing among the research community. As much as possible, data sharing requirements should not impede the work of researchers or cause undue financial or administrative burdens. To this end, we would like to see improvements in the administration of dbGaP to streamline its use and provide additional support for researchers who are utilizing this critical resource. Provision of robust federal funding for the National Institutes of Health will be critical for the Institute to continue providing investment in and improving the operation and widespread use of this important resource.

dbGaP Study Registration and Data Submission

Our research community has expressed a strong desire and willingness to share research data. However, it is our understanding that the user experience for those submitting data through dbGaP can be cumbersome and requires a lengthy registration and data submission process. We appreciate the commitment to ensuring that registered studies are appropriate and that data submission is standardized. However, we strongly encourage the provision of additional user support for researchers who are required to utilize the dbGaP program.

Additional support would be especially useful for researchers conducting large studies with data on multiple populations. There is often uncertainty around submitting complex datasets to dbGaP, and researchers may experience several failed submissions before they successfully upload their data. As researchers may only submit one dataset at a time, any stalls in submission can result in a backlog of data. Researchers are often still working through data submission requirements long after a study has been completed, and they are looking to move on to new projects and funding opportunities. We suggest providing personalized user support and additional guidance for submitting data from complex studies to help streamline this process and make it less burdensome on the research teams.

User support would also be helpful for smaller laboratories with limited staffing capacity. While submission to dbGaP is often mandatory for researchers receiving certain kinds of funding through NIH, there is no additional funding provided to be able to devote the time and resources to train an individual on the process for data submission. Rather than leave researchers to manage the learning curve alone, it would be immensely helpful for the NIH to provide additional personalized user support to train researchers who may be less experienced in submitting data through dbGaP or who may have less time to devote to this task because they manage multiple professional roles.

dbGaP Data Access Request (DAR) and Review

dbGaP provides much needed infrastructure for data storage and sharing, including the ability of a third party to monitor data use and enforce regulations against the misuse of data. We understand there is a difficult balance in facilitating efficient data sharing for researchers and maintaining strong data safety standards to protect patients who have donated their genetic information. While the process of submitting and reviewing data access requests may take additional time, we appreciate the strong review process required for accessing data in dbGaP and the NIH's commitment to ensuring that data are being used properly and patient data are well protected.

However, this process should be streamlined wherever possible, and we encourage the provision of additional resources in this area to ensure that each institute is able to provide a knowledgeable staffer to review data access requests. This would help prevent unnecessary delays and facilitate efficient use of time and research funding.

Policies for the Management and Use of dbGaP Data

We encourage the provision of additional time and resources to consider improvements for data collection and organization in dbGaP. Though we appreciate the need for standardized data collection, there is also a need to ensure that data submission is efficient and that the data available through dbGaP is a valuable resource for the research community at large. Streamlining the process for data submission and improving the user experience may also encourage external use of dbGaP by researchers who are not obligated to submit their data but who may be generating information that would be useful for other researchers.

dbGaP should also consider ways to collect and organize additional aggregate datasets that are convenient and attractive for researchers. While data submission for research studies usually requires individual genotype data, many researchers who are requesting to use data from dbGaP will be looking to aggregate the data prior to conducting their analyses. Giving researchers the option to additionally submit aggregate datasets that they already have available and making these data accessible to the research community would make dbGaP an even more usable and attractive resource.

Further, we appreciate this opportunity to provide comments on the data submission, management, and access processes for dbGaP, and we encourage ongoing dialogue and additional opportunities in the future to provide input on optimizing the operation and user experience of dbGaP.

*****

We would be happy to discuss any of these topics with you in greater detail.

Sincerely,

Preston W. Campbell, III, MD
President and Chief Executive Officer

Submission Date
04/06/2017
Name
UDN Coordinating Center
Primary Purpose of dbGaP Use
Study Registration / Data Submission
What is your level of experience with dbGaP?
Novice (used dbGaP once, or only a few times)
Role/Other Role
Scientific Researcher
Type of Organization/Other Type
University
Name of Organization
Harvard Medical School/Undiagnosed Diseases Network (UDN)

Information Requested

1. dbGaP Study Registration and Data Submission
Submission documentation: Although there is extensive documentation for the dbGaP submission requirements, and although the data files were prepared and validated according to those requirements, the dbGaP staff introduced several new requirements late in our submission process. We suggest combining the submission guide, SRA guide, and other documents into a single, concise and complete document. Submission requirements: The requirements for phenotype data make it challenging to automate generation of the required spreadsheets. In particular, the need to having linking relatives in the pedigree file posed a significant challenge for our group. It was also hard to determine which variables were required and appropriate for our study. We suggest providing additional sets of example files for submitters. Submission status: We encourage more transparency about the status of samples and files after submission (e.g., status flags in the submission portal). Sequence of events: dbGaP staff do not allow submitters to test the genotype upload process until all phenotype files have been approved. This can lead to significant delays and impedes the ability of the submitter to interact directly with the dbGaP genotype group in the preparation of all of the needed files. Communication issues: Staff within the different dbGaP groups do not always appear to communicate effectively with each other with regards to the upload processes and do not always respond to submitter queries in a timely and effective manner. We suggest that to smooth the submission process both for the submitter and the dbGaP staff, the dbGaP curators and SRA staff work more closely together to collect and review the phenotype and genotype data, and work as a team to communicate with the submitter. Technical difficulties: Some technical difficulties have arisen during the upload process, most notably with the need to strip identifiers from the sequencing files and to replace them with new identifiers for deposit in dbGaP. Cohesion around automated submission: Our side of the process for submission was largely built around the capability of a user in our system to create the package to submit to dbGaP in an automated fashion. It wasn't initially clear to us that we could submit meta-data about the bam files through XML, nor retrieve the results of the submitted files programmatically. We also encountered different steps whenusing ascp to submit the BAM vs VCF data, the latter of which required a user's manual interaction with the dbGaP portal. Reviewing the submission process from both a one time user's perspective as well as continued automatic submissions may be helpful in isolating the areas that need more programmatic interfaces. Testing submissions: Expanding the test site to act as a full integration test of submitting data may be helpful for those submitting data via their own applications. Being able to upload a small file and getting feedback about its validity after transmission would make it easier to test submission applications.

2. dbGaP Data Access Request (DAR) and Review 3. Policies for the Management and Use of dbGaP Data

- **Alternate controlled-access models**
  Although we have not requested access to any studies, we have heard that the existing approval process is burdensome and discouraging for some investigators. We envision an alternative model where users can apply for 2 levels of access: Level 1 (registered access): users can register with an academic (.edu) email address to query a study data set for a specific set of phenotype characteristics to determine whether they would like to apply for controlled access. Level 2 (controlled access): full access to study data Level 2 access should include an improved way to browse data (i.e. ability to query by genomic position, etc.)
- **Benefits and risks associated with the availability of genomic study summary statistics**
- **NIH is interested in public comments on benefits and risks of such reference use of dbGaP data for research**

participants, patients, and the scientific community

4. General Comments. NIH welcomes general comments on any other topics with regard to dbGaP data submission, access, and management

Thank you for the opportunity to submit comments on a valuable and important national resource.

Submission Date
04/07/2017
Name
Luis Serrano
Primary Purpose of dbGaP Use
Data Access / Download
What is your level of experience with dbGaP?
Role/Other Role
Institutional Official
Type of Organization/Other Type
Nonprofit Research Organization
Name of Organization
Center for Genomic Regulation (CRG)

Information Requested

1. dbGaP Study Registration and Data Submission

2. dbGaP Data Access Request (DAR) and Review
There are still areas that could be improved/simplified:
- When we recruited a new IT-director at CRG, it took a month to have his details included in the system (see NCBI tracking system #16792871). Initially, dbGaP Helpdesk asked: "For each approved dbGaP project, the PI should update their application. I have appended instructions on how to do this. You can pass these instructions along to the PIs." We have over 10 different PIs accessing dbGaP databases, and it would have been a real mess to get them to change the IT Director for CRG by themselves. After insisting that this was a purely administrative issue, the dbGaP helpdesk could make this change for all the projects involving CRG. We believe that this should be the default option.
- Similarly, changing the roles within a DAR should be simplified. It happens that a PI of a project is leaving to another institute, and that he/she is replaced by another senior researcher. At present, there is no straightforward way to do this.

3. Policies for the Management and Use of dbGaP Data

- **Alternate controlled-access models**
- **Benefits and risks associated with the availability of genomic study summary statistics**
- **NIH is interested in public comments on benefits and risks of such reference use of dbGaP data for research participants, patients, and the scientific community**

4. General Comments. NIH welcomes general comments on any other topics with regard to dbGaP data submission, access, and management

Submission Date
04/07/2017
Name
S.P.A. Drummond,PhD, President
Primary Purpose of dbGaP Use
Both Study Registration / Data Submission and Data Access / Download
What is your level of experience with dbGaP?
Experienced (used many times over the course of several years)
Role/Other Role
SRS President
Type of Organization/Other Type
Professional Org/Association
Name of Organization
Sleep Research Society

Information Requested

1. dbGaP Study Registration and Data Submission
We laud NIH efforts to improve processes for data submission, access and management for the Genotypes and Phenotypes (dbGAP) controlled-access data repository that serves as a platform for genomic data sharing. The multi-step dbGAP Study Registration process is navigable but complicated, and data submission of individual level phenotype data can on occasion be a lengthy process, necessitating multiple iterations of data formatting and quality control over many months. A more streamlined process with more effective communication between the submitter and dbGAP to clarify specific technical aspects of data formatting would be desirable. The dbGAP Study Registration process could be simplified for rapid submission of summary analysis results only for datasets without individual level data. The current process is sufficiently burdensome that researchers in our community have opted to deposit aggregate summary statistics for publications at public consortia or individual websites rather than in the more secure and central dbGAP repository.

2. dbGaP Data Access Request (DAR) and Review
The dbGAP Data Access Request process is cumbersome and non-intuitive for those not familiar with the initial requirements from the PI, institutional IT and signing officials, and IRB approvals. With some practice, requests for access or renewal can be rapidly submitted, with a quick turnaround time for review. Clear, easy-to-access step by step instructions for submitting requests would aid those unfamiliar with the system and increase utilization of dbGAP. Similarly, data download is not easy, with limitations such as inadequate search tools to identify subsets of study data to download and poorly described download and decryption software and processes. More advanced search features and step by step instructions here would also likely encourage more investigators to utilize dbGAP.

3. Policies for the Management and Use of dbGaP Data

- **Alternate controlled-access models**
  We consider sharing of genomic study summary statistics in a central repository a powerful means to catalyze research in genetics and genomics of sleep and circadian rhythms and their related disorders. Benefits include higher likelihood of use in secondary analyses by the research community – e.g. for replication, meta-analysis and integrative genomics including cross-phenotype, gene-based, pathway- based or global analyses of genetic architecture – that may help to identify functional variants and illuminate genetic links to other biological systems, and less reliance on individual level data access or duplication of work. We advocate the provision of unrestricted access or a simple controlled access process to obtain genomic summary statistics given concerns of identifiability of an individual participant from summary statistics, participant privacy and protection of vulnerable populations are appropriately addressed. Furthermore, we support new policy regarding use of controlled-access dbGAP data on approved cloud computing platforms and would like NIH to consider such platforms to enable secure collaborative research across institutions using dbGAP data.
- **Benefits and risks associated with the availability of genomic study summary statistics**
- **NIH is interested in public comments on benefits and risks of such reference use of dbGaP data for research participants, patients, and the scientific community**

4. General Comments. NIH welcomes general comments on any other topics with regard to dbGaP data submission, access, and management

The Sleep Research Society encourages data sharing and deposit of individual level genomic and phenotype data in the controlled-access dbGAP data repository by sleep researchers. The SRS advocates use of dbGAP datasets in secondary analyses by the sleep research community to enhance the pace of scientific research in sleep genomics, with appropriate attribution of credit and/or collaboration with primary studies. The SRS urges dbGAP to simplify data submission, access, download and collaborative sharing processes, to enhance the ease of use of dbGAP and welcomes the inclusion of approved cloud computing platforms. Further, the SRS encourages a public or a simple controlled-access searchable platform for genomic summary statistics. If some or all of these suggestions are accommodated, the SRS will proactively promote the availability of dbGAP to our members and encourage them to considering contributing to and accessing dbGAP to further utilization of this valuable resource.

Additional Document

Sleep
Research
Society®

April 6, 2017

Office of Science Policy (OSP)
National Institutes of Health
6705 Rockledge Drive, Suite 750
Bethesda, MD 20892

Re: Sleep Research Society reply to NIH Request for Information on Processes for database of Genotypes and Phenotypes (dbGaP) Data Submission, Access, and Management

The Sleep Research Society is an organization for scientific investigators who conduct research and provide education in sleep, circadian rhythms and their related disorders. The SRS mission is to foster scientific investigation on all aspects of sleep and its disorders, to promote training and education in sleep research and to provide forums for the exchange of knowledge pertaining to sleep.

We consider responsible sharing of relevant individual-level and/or aggregate phenotype and genomic data a key priority and opportunity to accelerate sleep and circadian rhythms genomic research across and beyond the sleep research community.

1. dbGAP Study Registration and Data Submission
We laud NIH efforts to improve processes for data submission, access and management for the Genotypes and Phenotypes (dbGAP) controlled-access data repository that serves as a platform for genomic data sharing.

The multi-step dbGAP Study Registration process is navigable but complicated, and data submission of individual level phenotype data can on occasion be a lengthy process, necessitating multiple iterations of data formatting and quality control over many months. A more streamlined process with more effective communication between the submitter and dbGAP to clarify specific technical aspects of data formatting would be desirable.

The dbGAP Study Registration process could be simplified for rapid submission of summary analysis results only for datasets without individual level data. The current process is sufficiently burdensome that researchers in our community have opted to deposit aggregate summary statistics for publications at public consortia or individual websites rather than in the more secure and central dbGAP repository.

2. dbGAP Data Access Request and Review
The dbGAP Data Access Request process is cumbersome and non-intuitive for those not familiar with the initial requirements from the PI, institutional IT and

signing officials, and IRB approvals. With some practice, requests for access or renewal can be rapidly submitted, with a quick turnaround time for review. Clear, easy-to-access step by step instructions for submitting requests would aid those unfamiliar with the system and increase utilization of dbGAP. Similarly, data download is not easy, with limitations such as inadequate search tools to identify subsets of study data to download and poorly described download and decryption software and processes. More advanced search features and step by step instructions here would also likely encourage more investigators to utilize dbGAP.

3. Policies for the Management and Use of dbGAP Data

We consider sharing of genomic study summary statistics in a central repository a powerful means to catalyze research in genetics and genomics of sleep and circadian rhythms and their related disorders. Benefits include higher likelihood of use in secondary analyses by the research community – e.g. for replication, meta-analysis and integrative genomics including cross-phenotype, gene-based, pathway-based or global analyses of genetic architecture – that may help to identify functional variants and illuminate genetic links to other biological systems, and less reliance on individual level data access or duplication of work. We advocate the provision of unrestricted access or a simple controlled access process to obtain genomic summary statistics given concerns of identifiability of an individual participant from summary statistics, participant privacy and protection of vulnerable populations are appropriately addressed.

Furthermore, we support new policy regarding use of controlled-access dbGAP data on approved cloud computing platforms and would like NIH to consider such platforms to enable secure collaborative research across institutions using dbGAP data.

4. General Comments

The Sleep Research Society encourages data sharing and deposit of individual level genomic and phenotype data in the controlled-access dbGAP data repository by sleep researchers. The SRS advocates use of dbGAP datasets in secondary analyses by the sleep research community to enhance the pace of scientific research in sleep genomics, with appropriate attribution of credit and/or collaboration with primary studies. The SRS urges dbGAP to simplify data submission, access, download and collaborative sharing processes, to enhance the ease of use of dbGAP and welcomes the inclusion of approved cloud computing platforms. Further, the SRS encourages a public or a simple controlled-access searchable platform for genomic summary statistics. If some or all of these suggestions are accommodated, the SRS will proactively promote the availability of dbGAP to our members and encourage them to considering contributing to and accessing dbGAP to further utilization of this valuable resource.

Sincerely,

Sean Drummond PhD

Sean Drummond, Ph.D.
President
Sleep Research Society

Richa Saxena

Richa Saxena, PhD
Associate Professor of Anesthesia
Center for Genomic Medicine
Massachusetts General Hospital, Harvard Medical School

Submission Date
04/07/2017
Name
Li-San Wang
Primary Purpose of dbGaP Use
Both Study Registration / Data Submission and Data Access / Download
What is your level of experience with dbGaP?
Experienced (used many times over the course of several years)
Role/Other Role
Project data coordinator
Type of Organization/Other Type
University
Name of Organization
University of Pennsylvania

Information Requested

1. dbGaP Study Registration and Data Submission
We have no feedback at this time.

2. dbGaP Data Access Request (DAR) and Review
Through our experience working with several DACs, we have some suggestions regarding the DAC system: 1. Each DAC seems to act independently, without consistent turnaround time or consistent consideration of application content. A more uniform guideline on how decisions are made would reduce confusion and improve review efficiency. 2. If a Data Access Request (DAR) is rejected, there is no appeal process. We recommend a system by which a rejected applicant can appeal the rejection to another DAC, or a panel of individuals who are not on the applicant's DAC.

3. Policies for the Management and Use of dbGaP Data

- **Alternate controlled-access models**
  1. Regarding data management and distribution, we recommend a decentralized or hybrid model for data sharing. a. A centralized model as of now make the process inefficient: i. The data sharing process becomes inflexible since there has to be a single process that handles every project. This means the process has to implement the most stringent conditions to satisfy all project needs. ii. dbGaP/SRA has to handle all projects using the same system and study templates; difficult for individual studies to manage how their datasets are summarized and presented. iii. Turnaround time from submitting data to releasing data to the public is longer since there is only one queue for data processing. iv. There is currently no rule regarding return of derived data. This cannot be easily implemented just by dbGaP without help from all the communities. b. Investigators and institutions closely involved with generating data from each project is more familiar with the data, the best ways to share the data, and the best ways to use the data. This also opens new opportunities for data dissemination; for example, each project can develop its own databases based on analysis results and summary statistics from its own datasets. c. We recommend a hybrid model i. NIH and dbGaP develop a centralized catalog as it is now that registers all studies ii. Individual projects can decide to deposit data into dbGaP for sharing or develop their own data sharing initiatives abiding by rules set by NIH iii. NIH and dbGaP develops documents, guidelines, and information services to facilitate data sharing by individual studies. We expect this will only be done for very large projects, which will benefit from the additional flexibility if they are allowed to set up their own data repositories. 2. GDS Consent Levels a. The current way for certifying consent levels for subjects recruited/data generated before January 2015 (the time when Genome Data Sharing Policy – GDS – was enacted) is very flexible, but it also leads to difficulties in practice. b. Individual institution IRBs may choose to always fill in their own consent levels on their GDS Institutional Certification form. It is a logistic nightmare for both data managers and requestors to have to deal with more than 10 (potentially hundreds of) different consent levels. c. It will greatly facilitate the process if OSP could develop additional guidelines to reduce the number of different consent levels whenever possible.
- **Benefits and risks associated with the availability of genomic study summary statistics**
  Genome-wide summary statistics. We recommend general use under a qualified access process for genome-wide summary statistics based on the following reasons: a. There is non-negligible risk for identifiability with genome-wide summary statistics, although such risk is probably very low in practice and no one has demonstrated it is always doable.

Therefore a qualified access process ensures only trained qualified investigators have access to these data and they will not violate policies related to human subjects data use. b. Genome-wide summary statistics are aggregate statistics and it is extremely hard to generate multiple versions if the participants fall under multiple consent groups. Therefore we recommend general use to reduce the submitter's burden and maximize data sharing. Currently dbGaP only shares genome-wide summary statistics if the individual data of every subject used for the statistics is deposited into dbGaP for sharing. We recommend relaxing this restriction for two reasons: (1) it may be difficult when a meta-analysis includes cohorts that cannot be deposited into dbGaP (e.g. studies from abroad); (2) based on the arguments above on the low risk of identification and intent to maximize data use, the benefit of sharing summary statistics widely and early greatly outweighs the risk.

- **NIH is interested in public comments on benefits and risks of such reference use of dbGaP data for research participants, patients, and the scientific community**
  We have no feedback at this time.

4. General Comments. NIH welcomes general comments on any other topics with regard to dbGaP data submission, access, and management

Submission Date
04/07/2017
Name
Liam Curren
Primary Purpose of dbGaP Use
Data Access / Download
What is your level of experience with dbGaP?
Experienced (used many times over the course of several years)
Role/Other Role
Institutional Official
Type of Organization/Other Type
For-profit R&D Company
Name of Organization
Genomics plc

Information Requested

1. dbGaP Study Registration and Data Submission
N/A
2. dbGaP Data Access Request (DAR) and Review
We have found Data Access Committee (DAC) staff to be helpful, responsive, and professional when we have needed to discuss matters relating to our project, be it over email or phone. They have provided us with clear advice, which has helped us to access data from studies that we are interested in. Downloading data from dbGaP is straightforward, however we have seen many examples of confusingly structured study files. Common problems include: (i) a lack of standardised format for data types; (ii) large numbers of data files with sparse instructions on how they should be used; (iii) a lack of clear instructions that would help us distinguish between - and use - files submitted by users and files that have been generated by dbGaP; and (iv) a tendency to use nested tarballs, with many confusing levels of compressed files. These all contribute to making the processing of individual-level data files often very challenging. It would also be helpful if a primary paper citation was included within the dbGaP study information. We periodically request data from additional studies under the scope of our existing RUS. Each time we submit such a request, the dbGaP system automatically re-applies to all the DACs in respect of the studies for which we have already received approvals. We assume that this requires DACs to conduct a repeat review and approval of our request, and are wary of the unnecessary burden this creates for the DACs. It would be more efficient if additional requests for data, made under the same RUS, were processed in isolation. We have faced occasional problems when some DACs have requested that we expressly mention in our RUS that we agree to abide by specific obligations contained in the relevant DUCs. This uses up some text in the RUS - which is already subject to a rather tight word limit - and does seem somewhat unnecessary: our Signing Official will already have formally agreed to comply with all the obligations contained in the relevant DUCs. The project renewal process have been straightforward and easy to manage.

3. Policies for the Management and Use of dbGaP Data

- **Alternate controlled-access models**
  N/A
- **Benefits and risks associated with the availability of genomic study summary statistics**
  We believe that a controlled-access route for summary statistics in dbGaP, and other repositories, is burdensome and most likely unnecessary. There is a growing practice within the genomics research sector for summary statistics to be made - relatively - freely available e.g. without need for detailed application and review. A lower-effort route to good quality summary statistics would in our opinion be beneficial for all researchers. The privacy risks associated with this type of data are minimal, however we can appreciate the need for some accountability of data users. A simple registration system on dbGaP, accompanied by straightforward terms of access (e.g. use only in bona fide scientific research in academia, clinics, or industry; attribution/acknowledgement for the original study authors; plus a prohibition on attempting re-identification of research participants) would ensure that more summary statistics were utilised in new research. That could only be a good thing. Unfortunately, much of the summary statistic data made available in the analysis files on the dbGaP FTP site appears to be in a rather poor state. Common problems we have encountered include: (i) betas without a +/- sign; (ii) betas bearing no relation to the A1/A2 alleles in the same file; (iii) inconsistent content in header description information; (iv) unclear communication of important information (e.g.

sample size, analysis method are often not apparent); (v) a lack of ‘missing information’ values in data files; and (vi) entirely empty columns in some data files. All of these combine to make analysis of the summary statistic data very time-consuming. We appreciate that dbGaP is not resourced to conduct a check of all data submitted to it, however we would encourage a clearer requirement - or at the very least, a recommendation - for providers of summary statistics to include the following content: (i) both alleles; (ii) effect beta or odds ratio; (iii) effect SE or upper/lower bounds of odds ratio; (iv) P-value; and (v) rsID and/or chromosome/position. Finally, including information about which allele an effect pertains to is critical to the utility of summary statistics data. This can easily be done by, for example, labelling allele columns ‘effect_allele’ and ‘other_allele’, or by having accompanying information in a README file.

- **NIH is interested in public comments on benefits and risks of such reference use of dbGaP data for research participants, patients, and the scientific community**
  We are broadly supportive of the wider use of genetic information for clinical reference uses. However, in the absence of enforcement of data standardisation by most data repositories, it is very important that the quality of the data proposed to be used is checked carefully by end users.

4. General Comments. NIH welcomes general comments on any other topics with regard to dbGaP data submission, access, and management
dbGaP is an incredibly valuable resource for the research community, and we are grateful for the significant efforts made by the NIH staff to maintain and develop it. We would support reducing the administrative burden for accessing summary statistics data on dbGaP. At the same time, we would strongly encourage clearer guidance and - if resources allowed - checking of the quality and completeness of data, particularly summary statistics, hosted on dbGaP.

Submission Date
04/07/2017
Name
P. Pearl O'Rourke
Primary Purpose of dbGaP Use
Study Registration / Data Submission
What is your level of experience with dbGaP?
Experienced (used many times over the course of several years)
Role/Other Role
Institutional Official
Type of Organization/Other Type
Health Care Delivery Organization
Name of Organization
Partners Health Care

Information Requested

1. dbGaP Study Registration and Data Submission

2. dbGaP Data Access Request (DAR) and Review

3. Policies for the Management and Use of dbGaP Data

- **Alternate controlled-access models**
  In response to the suggestion for alternate controlled access models, the Partners HRA supports the use of a registered access system for use of genomic summary statistics. A registered access system would allow for automated data queries based on predetermined cohort criteria categorized by incoming data use limitations. The system would include clear eligibility criteria regarding who could register and a process to confirm registrants who meet these standards. We also assume that a data use agreement would be required and an auditable tracking system would be in place. With such a system, appropriate users or use cases could be tracked.
- **Benefits and risks associated with the availability of genomic study summary statistics**
  We recognize that access to genomic summary statistics is an important resource for health/biomedical research and the risk of disclosing this data is low: Partners HRA supports the management of most genomic summary data in a controlled or open access system. But, data use limitations should be available if and when the risk of summary statistics is elevated, for example when the data includes vulnerable populations or sensitive information. In addition, it would be helpful to have guidance regarding definition of any categories of data that have virtually no risk for re-identification, for example mRNA, certain transcriptome analyses accompanied by minimal phenotypic data. If such categories can be defined, we suggest that NIH re-evaluate the current Certification process and specifically the role of the Institutional Review Board (IRB) in that process.
- **NIH is interested in public comments on benefits and risks of such reference use of dbGaP data for research participants, patients, and the scientific community**

4. General Comments. NIH welcomes general comments on any other topics with regard to dbGaP data submission, access, and management

Pre-award Institutional Certification and the Just in Time (JIT) process: At JIT notification, many researchers haven't submitted an IRB protocol. Conducting the analysis for Certification, comparing data sharing plans for consistency with the GDS policy is difficult without a protocol. Investigators may rush to submit a protocol –these are often incomplete requiring multiple rounds of communication, rendering the process inefficient. Without a protocol, the IRB is asked to review the grant application. In these cases, a "Provisional Certification" can be provided with promise for future
Institutional Certification. The Provisional Certification "assures" that future data sharing outlined in the genomic data sharing plan will be consistent with the elements in an Institutional Certification. But grant submissions are not equivalent to IRB submissions and making these assurances in the absence of a formal IRB protocol can be challenging. Additionally, there is no guidance on the NIH GDS website for institutions regarding how and when to use the Provisional Certification process and how soon a final Institutional Certification must be submitted. Guidance material regarding the use of Provisional Certifications would be helpful. The Provisional Certification template has not been accepted by all NIH Institutes. We suggest a standardized

process across NIH. If Pre-Award Certification is necessary, we suggest NIH and institutions require researchers to initiate the certification process before the JIT period during the first favorable scoring cycle. B. We propose the following changes for the certification: • Allow limited institutional customization and design certification forms to accommodate such details. This could be an open text box designated for institutional revisions. o An example of institutional customization: the Pre-Award Certification letter notes compliance with all local, state and federal laws and consideration of risks of future data sharing to individuals and groups. Our institution has made the following changes to these assurances: • Certification is limited to compliance with local, Massachusetts (not all US states) & federal policies; • Given that decisions about subsequent data sharing are under the control of the DACs at NIH, it is not possible for local IRBs/Privacy boards to be able to identify all potential risks of subsequent sharing either to individuals or groups. Therefore we note that subsequent data sharing is reviewed by NIH DACs. o These changes are entered into the available text box found on page 2 of the form and labeled Data Use Limitation. This is not the most effective way to communicate our revisions. • We recommend an electronic portal for Institutional Certification assurances and sign off. This would be accessible by all parties involved – the researchers, NIH program office, IRB and Institutional Officer. This platform could mimic electronic protocol management systems whereby a researcher initiates the process and the certification is reviewed and signed by appropriate parties as needed. C. Additional support and education to better understand the implementation of: • the Data Use Limitation Modifiers (page 2 of IO Certification forms) • the public display of alleles/frequencies to NIH archives (i. e., dbSNP and dbVar) (bottom of page 1 of the Certification Form).

Submission Date
04/07/2017
Name
Sheila M Reynolds
Primary Purpose of dbGaP Use
Data Access / Download
What is your level of experience with dbGaP?
Experienced (used many times over the course of several years)
Role/Other Role
Scientific Researcher
Type of Organization/Other Type
Nonprofit Research Organization
Name of Organization
Institute for Systems Biology

Information Requested

1. dbGaP Study Registration and Data Submission
no comment, I am not familiar with data-submission

2. dbGaP Data Access Request (DAR) and Review
In my experience, dbGaP has been the place to request and obtain permission to access controlled data, but the data has always been somewhere else -- eg at the TCGA DCC, at CGHub, or at the GDC (since last year). So the notion of "downloading data from dbGaP" seems odd to me, although I have seen buttons and menus and such on the dbGaP website related to browsing and downloading data. I think that it would probably be helpful to clearly distinguish between DAR functions and where the data actually can be accessed. With the move towards the cloud, the data may also be available in multiple locations and a mechanism to be able to reference/find data in a variety of locations would be helpful -- eg if a user is looking for a particular WGS bam file, it would be useful to know if it exists in Google Cloud Storage, Amazon S3, etc in addition to at the "official repository" which might be the GDC in Chicago.

3. Policies for the Management and Use of dbGaP Data

- **Alternate controlled-access models**
  The ISB-CGC (NCI funded ISB Cancer Genomics Cloud pilot project) has pioneered a method by which users can be authenticated by a combination of their Google identity and their eRA Commons / NIH identity. Once the authentication process has succeeded, the user's authorization status is verified by checking the available project-specific dbGaP whitelist(s), and the user is then added to a "Google Group" which is on the ACL (access control list) for the controlled-access data that the user has requested. Once this process has completed (assuming all authorization and authentication checks were ok'd), the user has "direct" access to the data in the cloud bucket. We think that it is important that at this point the access to the data is through the native cloud APIs, and not throttled in any way by any additional layer which might slow down data access. Another significant aspect to the way the ISB-CGC provides and controls data access is to differentiate between "user credentials" and "service account credentials" -- generally in the cloud, VMs use "service account" credentials to access data rather than personal credentials since a service account can be granted more tightly controlled, "scoped-down" credentials. On the other hand, service accounts are generally accessible to all users who are members of a given "cloud project", so additional verification steps are necessary to make sure that all project members are also authorized to access the controlled data.
- **Benefits and risks associated with the availability of genomic study summary statistics**
  I think this has been a very significant issue for many many years and that being overly "conservative" on this front can really slow down the progress of research. Clear guidance from NIH on what type of data must be treated as "controlled" and at what point downstream analyses can be treated as "open" would benefit everyone. For example, right now, the GDC has reprocessed and re-called mutations on all of the TCGA data, but the open-access MAF files that are currently available were filtered with overly restrictive filters which have rendered these open-access MAFs pretty much useless. In some cases, statements have been made to the effect that "all results from all analyses based on controlled data are to be treated as controlled data" -- but that is clearly not the case. For example, in TCGA all RNA or DNA-derived information *starts* from controlled data but once it is "summarized" as gene- expression data or

copy-number segments, it is then treated as open-access. These are examples of data "aggregation" but for a single case/sample vs aggregating across a large cohort. Having everyone in the research community fearful of repercussions if they accidentally "leak" controlled-data is not beneficial for research. I believe that Joe Biden and the Cancer Moonshot have also expressed the view that making data easier to access is in everyone's interest -- including the cancer patients and those who have graciously shared their samples/data with researchers.

- **NIH is interested in public comments on benefits and risks of such reference use of dbGaP data for research participants, patients, and the scientific community**
  I strongly support the clinical use of genomic research data.

4. General Comments. NIH welcomes general comments on any other topics with regard to dbGaP data submission, access, and management

Submission Date
04/07/2017
Name
Rajni Samavedam
Primary Purpose of dbGaP Use
Browsing Unrestricted-Access Study Information
What is your level of experience with dbGaP?
Novice (used dbGaP once, or only a few times)
Role/Other Role
Role Scientific Consultant
Type of Organization/Other Type
Consulting Firm
Name of Organization
Booz Allen Hamilton

Information Requested

1. dbGaP Study Registration and Data Submission
Critical to the success of a data sharing platform like dbGaP is the number of studies it houses and makes available to investigators for secondary analysis. Therefore, study registration and data submission processes must be user friendly and efficient; however, current processes pose challenges. An NIH IC Program Official must initiate study registration, creating delays and difficulties for Principal Investigators (PIs), especially non-NIH PIs. Requiring an eRA Commons account is similarly restrictive without giving any advantage for account holders as they are not authorized to initiate submissions.
Obtaining an eRA account can take four weeks - a costly delay for investigators interested in submitting data linked to publications. These challenges discourage submissions to dbGAP – a submitter daunted by a year-long delay for submitting data from his Cell publication turned to the European Genome-phenome Archive to satisfy the journal's requirement. To address these challenges, dbGaP should consider additional methods for registering a submitter (institutional email/ORCID account, etc.), and facilitate submissions without requiring approval by NIH IC Officials, but requiring non-NIH submitters to link the submission to a peer reviewed publication. Not mandating the eRA account and NIH IC sponsoring requirements will improve efficiency and expand dbGaP considerably. If de-coupled from NIH IC, submitters from non-NIH institutions should bear the administrative costs for submission as a donation to NLM. The required Institutional Certification ensures consistency of data sharing with the participants' informed consent. However, obtaining Certification is problematic for completed studies where IRBs are no longer active. To circumvent this, dbGaP should consider informed consent review by the appropriate NIH IC's IRB, or the NIH OHSRP for non-NIH funded studies. Another consideration is to establish a central NIH ethics review board to streamline submission. Prospectively, the single IRB model proposed by the revised Common Rule will improve data sharing efficiency. Submitters also encounter technical challenges during data submission. Precision Medicine projects like NHLBI's TOPMed generate petabytes of data; recently, uploading files with 40,000 Whole Exome Sequences/Whole Genome Sequences was shut down due to bandwidth limitations. Even though NCBI recently set up an AWS cloud-based 10 GB DataConnect, higher levels of throughput is necessary for on-premise maintenance of dbGAP. A mirrored hybrid cloud environment for large throughput with elastic ingest capability would address the first and last mile issues. Implementing regional cloud availability zones, where data generated on the west coast can load into west coast instances will reduce ingestion bottleneck at Bethesda, followed by synchronizing the regional instances with on-premise dbGAP. Different research institutes generate data using different sequencing machines in varying data and metadata formats. To maximize reuse, dbGaP should utilize common data elements to enable data harmonization and standardization, similar to the Global Alliance for Genomics and Health community. To ensure better data discovery, access and citation, dbGaP should consider Digital Object Identifiers for all data assets. For distributed computing of large scale data, dbGaP should consider tools that allows maintenance of data structures in distributed dynamic and/or unreliable (e.g., user desktop) environments.

2. dbGaP Data Access Request (DAR) and Review
Similar to submission, the process for requesting access to individual-level dbGAP data requires an eRA Commons account and an NIH Program Official to initiate registration for data access for both NIH- and non-NIH funded researchers. While dbGaP provides access to summary-level data via the NIH Genome Browser tool, which is a welcomed step towards enabling discovery, access to this tool also requires an eRA login. This limits graduate students, post-doctoral fellows, and other young investigators to those affiliated with a PI who is an eRA account holder. dbGaP should provide free access to all researchers to browse and search for data, but require registration and approval by a Data Access Committee for access to data. However,

registration should be not be restricted to eRA account holders nor initiation by an NIH Program Official. Instead, as with submission, additional models for access such as through verified institutional email or ORCID account should be considered. These account holders will be required to obtain signatures from the PI of the proposed study and the institutional signing official in the Data Use Certification, thus coupling the provision of data only to investigators affiliated with a bona fide institution. This process will also eliminate the involvement of an NIH IC Program Official. Currently, data request approval decisions are made by the NIH IC that sponsors each study in dbGaP but given that our recommendation is to not have the submission tied to a specific NIH IC, we propose that dbGaP establish a process where approval is governed by dbGaP (for example, a Trans-NIH Data Access Committee), and if required, the data submitter-defined approving entity (for example, Study Steering Committee). This would eliminate having to establish and obtain approvals from individual Data Access Committees from each IC as is now the case. Additionally, the current review and approval process by the respective NIH IC Data Access Committee takes 1-2 months and dbGaP directs the requester to the Committee for any questions related to the request (both prior to and after making the request) – this is another major rate limiting step. Once a data request is approved, data are generally downloaded to a local server using Aspera connect. Although this is currently preferred as it enables aggregate analysis with private or other public data sets, as cloud- based storage and compute platforms, such as the NIH Commons, become more widespread, instead of downloading, the data could be analyzed in a common cloud environment. This would enable researchers to leverage cloud-scale storage and compute capacity. Download and upload is costly, compute is relatively less expensive, so this is another way to make the process more efficient for all.

3. Policies for the Management and Use of dbGaP Data

- **Alternate controlled-access models**
  To promote access of dbGaP data to the wider research community, beyond those with an eRA Commons account, alternate controlled-access models should be developed and implemented. One possible alternative is to enable users to establish data submission/access accounts with different "levels of trust", which would ensure security of data and privacy of participants through role-based access. The levels of trust may be based on institutional affiliation (emails), ORCID or trusted partners. This information would be included on the individual's account record and would be viewable by the public and dbGAP administrators. The "level of trust" could then be used to determine: 1. Process for data submission – automation could be used wherever possible for submitters with the highest level of trust (eRA Commons account holders or Trusted Partners). Other types of account holders may be able to submit studies but the submission process may require additional levels of review. 2. Levels of data access – individual-level data may be requested by eRA Commons account holders or Trusted Partners while others may be able to view summary-level data and/or metadata (either in part or in entirety). A controlled access model may also include requests from participants wanting to access and analyze their own data during the embargo period. Public access to higher-level, de-identified data could be provided to all levels of account holders to expand access across the global scientific community to stimulate hypothesis generation.
- **Benefits and risks associated with the availability of genomic study summary statistics**
  Genomic study summary statistics have tremendous scientific value for secondary analyses, including meta-analysis, joint and conditional multi-SNP analysis, and polygenic disease risk prediction. Generally, the results approximate what can be achieved using the raw genotypes, with theoretical bounds on the error. Analyses such as disease risk prediction become clinically useful as training set sizes increase, thus necessitating broad sharing of summary statistics to maximize benefits. Summary statistics are in general of lower risk than individual genotypes. However, the presence of a single individual's genotype can bias the aggregate statistics in a detectable way. One potential risk is the "attribute disclosure attack from DNA", where an attacker uses a person's genotype along with summary statistics to determine whether that person was in the case or control group. However, this requires the attacker to have the genotype of the target individual and prior knowledge that the individual was a study participant. Some measures for mitigating this risk include:
  (1) restricting access to the individual's genotype,
  (2) excluding participant metadata that increases the attacker's prior knowledge that the targeted individual was a participant, and (3) increasing study sample size to minimize risk to any individual participant. Vulnerable populations pose another risk - for many secondary analyses, it is necessary to use linkage disequilibrium information computed from a reference panel (e.g. 1000 genomes, UK10K) that is sufficiently similar to the target sample. However, many minority groups remain under-represented in large genome samples, potentially generating inaccurate results for those populations. This highlights the need to increase data submissions from studies on minority populations and rare diseases, to increase their chance of discovery. One alternative to the controlled access model is providing researchers with an API for querying summary statistics and recording those queries, which could be audited in the event of a privacy attack. Another option is to perform all computations in the cloud with sensitive data never leaving the cloud

environment, similar to the NCI's Cancer Genomic Cloud pilots or the Million Veteran Program's Genomic Information System for Integrative Science, which Booz Allen helped implement. Potential approaches to mitigate risks associated with access include:
(1) A dynamic consent model where presence in publicly available summary statistics is consented to but later withdrawn if desired. Given a full picture of the risks and benefits, many participants may be willing to consent to publicly sharing summary statistics. Moreover, the risk can be mitigated by evaluating the sensitivity of the trait being tested. Many GWASs test physical attributes (e.g., height, weight) that are not sensitive in nature and could be shared more openly.
(2) A computational technique called differential privacy, used by major tech companies to protect user data. Differential privacy transforms the data such that it is effectively the same regardless of the presence or absence of a single individual. Some studies of applying differential privacy to summary statistics have found that it adds an unacceptable level of noise; however, new techniques have been developed that improve its performance for GWAS.

- **NIH is interested in public comments on benefits and risks of such reference use of dbGaP data for research participants, patients, and the scientific community**
  As the NIH's central portal for human genomic, phenotypic, and exposure datasets, dbGAP has the potential to stimulate significant discovery and aid in the translation of scientific research data into the clinic. This is particularly relevant to the Precision Medicine initiative, which has the stated goal of revolutionizing healthcare outcomes by tailoring therapeutic regimens according to an individual's genome. It is important to note, however, that the dbGAP data are fundamental research, and discoveries using this resource will require clinical validation. However, the true value of the data in translating meaning of the human genetic code will only be understood when a critical threshold of data are collected and made easily accessible to the research community. Information from well-designed human subject research deposited into dbGaP could elevate or lower the level of evidence for a molecular marker contributing toward a particular disease or treatment response. Consistency in the direction and magnitude of an association across diverse populations and molecular phenotypes would provide medical geneticists with strong support for a marker-disease or marker-treatment response relationship. Informed decisions could then be made in devising health management plans that benefit patients with a genetic marker that are at-risk or living with an existing associated condition. Although data in dbGaP are understandably skewed toward the most common or well-funded conditions in research, searches for a rare disease gene locus may uncover relationships with common traits that explain some symptoms and/or indicate the potential for off-label drug use, as well as stimulate new hypotheses for research. The capability of identifying such relationships will continue to grow as more and more data are deposited. While reference use of dbGaP would benefit patients and the scientific community, several limitations exist that could lead to the risk of misinterpreting data leading to harmful clinical decisions. Medical geneticists will need to clearly distinguish between data generated using clinically validated assays vs other assays that have not undergone the same level of quality testing and assurance, and would thus impact confidence in the results. A sufficient understanding of study design characteristics, materials, and assay utility is necessary to appropriately interpret dbGaP results such as the tissue-specificity of epigenetic and gene expression results, timing of biospecimen collection in relation to disease, or oversampling of individuals with a specific disease subtype into a study. Users also must understand that the analysis of high-dimensional data are riddled with false positives and necessitate replication in external datasets in order to make appropriate conclusions. Even with sufficient knowledge to interpret the evidence, one potential drawback to the use of dbGaP as reference data is the different thresholds used by one medical geneticist vs another, leading to treatment decisions that vary between hospitals and providers. Alternatively, treatment decisions based on dbGaP data may become widely accepted in standard practice without sufficient testing of clinical validity and utility in randomized trials. Finally, the wide-spread use of dbGaP data for clinical reference poses a potential risk to research participants if data security practices are inadequate.

4. General Comments. NIH welcomes general comments on any other topics with regard to dbGaP data submission, access, and management

To further maximize the utility and efficiency, dbGaP might want to consider the following: - Enhance the 'look and feel' of dbGaP website and implement a user friendly dynamic interface so that those who are first time users can easily navigate to the desired pages. - Provide an easily locatable

User Instruction Manual/How-To Guide that is detailed and interactive on the homepage. Currently, there is no one place to get all the information on the policies, processes and instructions on the website or the homepage – the most succinct description on dbGaP is in a publication from 2015 when dbGaP was established (Nat Genet. 2007 Oct; 39(10): 1181–1186) and not on the dbGaP website itself. - Increase meaningful reuse of the data by making it FAIR
– Findable, Accessible, Interoperable, and Reusable - Establish interoperability with other large and commonly used NIH (first) and external repositories (later) to enable data integration and analysis - Role of performance metrics: Data access metrics are useful, but emphasis should also be on the stories behind the numbers – especially studies that laid the foundation for significant findings or breakthroughs; Reaching out to users to assess needs and facilitate submission and request processes; metrics related to ease of submission and request will be useful for adjusting implementation - Develop tools and upgrade IT infrastructure to automate data submission and data access; current process requires significant manual curation. Although a fully automated submission and access process may not be possible, upgrading IT infrastructure and developing tools for automation wherever possible would dramatically improve the efficiency of processes. This may require initial investment but the net effect would likely result in a cost savings due to reduction in curation over the coming decades. - Enhance collaboration and analytics – Meaningful GxP (Genotype by Phenotype) associations and data for possible clinically relevant variants (e.g. GWAS LOD score > 3) essentially is not made available rapidly enough for multidisciplinary, multi institute collaborative consortium projects. This is a major disservice to the scientific community and runs counter to the NIH data sharing policies put in place from the original Human Genome Project and the 1000 Genome project which both aimed to release data to the public as soon as reasonably possible. dbGaP might consider various strategies such as allowing researchers to link data sets together with other data sets in their NIH Commons account or fostering sharing of analytic pipelines across data sets to facilitate reproducible research. - Ensure that molecular variant data generated by clinical labs on limited patient samples and the associated “clinical” interpretations are maintained separately in the database until there are sufficient sample size from pooled data on specific variants available to reliably conclude significance. - Represent clinically actionable subsets of genomic data in easily accessible ways for use by the clinical community.

Submission Date
04/07/2017
Name
Wayne Huggins
Primary Purpose of dbGaP Use
Data Access / Download
What is your level of experience with dbGaP?
Experienced (used many times over the course of several years)
Role/Other Role
Scientific Researcher
Type of Organization/Other Type
Nonprofit Research Organization
Name of Organization
RTI International

Information Requested

1. dbGaP Study Registration and Data Submission

2. dbGaP Data Access Request (DAR) and Review
I would suggest functionality that makes it possible to add collaborators from other institutions to a dbGaP data access request
• It would be optimal to be able to see the meeting schedules for individual Data Access Committees to get a sense of the timeline for data access request review
• I would suggest an option to add studies, variables, and datasets to a cart or basket while browsing. I often find myself navigating between two copies of dbGaP in different browsers: one for browsing individual studies or variables and another for either selecting those variable datasets from those studies in the file selector for download. It would be great to be able to add items to a cart and then have ready access to just those items from the download page.
• The IRB requirement for access to some of the studies seems a bit excessive

3. Policies for the Management and Use of dbGaP Data

- **Alternate controlled-access models**
- **Benefits and risks associated with the availability of genomic study summary statistics**
- **NIH is interested in public comments on benefits and risks of such reference use of dbGaP data for research participants, patients, and the scientific community**

4. General Comments. NIH welcomes general comments on any other topics with regard to dbGaP data submission, access, and management

Submission Date
04/07/2017
Name
Stacey Donnelly
Primary Purpose of dbGaP Use
Both Study Registration / Data Submission and Data Access / Download
What is your level of experience with dbGaP?
Experienced (used many times over the course of several years)
Role/Other Role
Institutional Official
Type of Organization/Other Type
Nonprofit Research Organization
Name of Organization
The Broad Institute, Inc

Information Requested

1. dbGaP Study Registration and Data Submission
Problem: The study registration and submission process is overly complex, manual, lacks a systematic manner in which data is curated, and does not accurately capture critical phenotypic information. Solution: 1. Automate sample registration process. Much of the sample registration process is manual and overly time consuming. Currently, this registration requires communicating directly with a specific dbGaP officer, which is not scalable. 2. Develop simple tools for the deposition of phenotypic data. Phenotypes are not structured on deposition, greatly complicating meaningful secondary use. NIH must either develop or encourage others to develop tools for structuring phenotypes. These structured phenotypes must be part of a comprehensive ontology system to enable flexible queries and searching. 3. Require deposition of all phenotypic data that is published. Again, meaningful secondary use requires good phenotypic data. If investigators are funded by NIH and have the phenotypic data and the tools are provided for easy deposition (#2 above), NIH should require the deposition of the data (at minimal, the phenotypic data that was already included in a publication). 4. Data Submission and Curation.
There exist several groups with expertise in large-scale data engineering, as a result of efforts such as TCGA, GTEx, ExAC and TOPMed. This expertise should be leveraged by creating a small number of centers (but more than one), that are charged with a. Processing of genomic data with various best practices pipelines. This mandate should include reprocessing data with new builds of the genome, or when there is an algorithmic advance in variant calling significant enough to justify it. b. Mapping phenotypes to ontologies. c. Structuring data use restrictions so that they are mapped to a standard ontology (see Section 2, below). 5. Computing Infrastructure: The curated datasets should be placed onto public clouds for use by the community (more than one is crucial to maintain competition), in addition to NIH-owned servers from which it can be downloaded. It is crucial that these datasets be directly accessible by users, and that users not be required to go through any given platform to access a given dataset. 6. Access Control: dbGaP should continue to adjudicate access to publicly funded datasets. Under this model, dbGaP would continue to be an identity provider that adjudicates who is a "trusted researcher," and researchers would continue to apply to this body to receive permissions to access various datasets. The data should be stored on one or more cloud infrastructures; as access is granted, this would be mirrored across infrastructures. dbGaP should make whitelists of which researchers are allowed to access the various applications publicly available via APIs so that third-party applications could respect these permissions. In summary, the above model would facilitate access: a. By ceding data curation to groups with expertise in it, this work would no longer fall squarely on the shoulders of dbGaP. b. By making data use restrictions machine-readable, the work of DACs would be greatly simplified. c. By utilizing clouds, there would be less need for downloading data to new environments, which significantly taxes the current system.

2. dbGaP Data Access Request (DAR) and Review
Problem: The system for granting data access is entirely manual. This process is both: (a) highly variable based on the Data Access Committee; (b) simply not scalable. Additionally, the annual renewal and reporting process is extraordinarily burdensome for investigators to complete and NIH staff to review and provides little benefit. Solutions: I. Standardize Data Use Letters (DULs) and Data Access Requests (DARs). The Broad has developed a user friendly tool and has piloted this experiment. We inspected a collection of nearly 132 DULs at the Broad. From that exercise, we determined that 95% could be structured into an ontology that contained the following 5 main categories (see Figure below):

i) disease-specific restrictions,
ii) ii) commercial restrictions,

iii) restrictions to special populations, iv) restrictions on research use, v) General Research Use. If DULs were structured to follow a standard ontology, researchers would be able to search for datasets consistent with their research purpose. It would also greatly facilitate the work of the DAC, as the effort of checking whether a DAR is consistent with a DUL would be automatable and thus scalable. Towards this end, our Institute has developed an open-source software package ("DUOS") for structuring DULs and filtering them by research purpose. This software package also contains interfaces for DACs to facilitate review of data access requests. We would happily donate this software to dbGaP for use by its DACs. II. Phenotype-based search Just as the previous example noted the need to data-use-enabled search, there is a significant need for phenotype-based search. Researchers would greatly benefit from the ability to search for all samples with a given phenotype in constructing cohorts. This search should be ontology-aware (e.g., search for "cancer" includes "angiosarcoma" etc'). III. Extending and Renewing Data Access Requests Currently, if a researcher has approval to study one dataset and later wants to utilize a newly deposited dataset for the same research, this entails an entirely new application. There should be a mechanism by which researchers can extend an already-approved application to include additional datasets as they emerge. Finally, the annual renewal process is unnecessarily burdensome for both Investigators and NIH staff. We propose that all investigators agree to a code of conduct rather than annual renewal and that an audit function by the "trusted partner" replace annual renewals. IV. Better coordination of DARs with Cohorts Many of the traditional large-scale cohorts are governed by an added layer of access control--either a local IRB or a DAC that is affiliated with the cohort. This can greatly complicate and prolong the process of accessing data, and we have experienced inconsistency between cohorts with regard to protocols and policies. We propose that this additional layer be streamlined where one representative from each cohort is designated to make decisions in a time-limited fashion. And, if there are additional limitation on data use, the group should make them entirely transparent so that researchers will only ask for the data if they can meet any of the additional layers of review.

3. Policies for the Management and Use of dbGaP Data

- **Alternate controlled-access models**

  Problem 1: 'Bring data to the researchers'- researchers are forced to make copies of the data. dbGaP presumes that researchers will download data to their own infrastructure. Accessibility is a significant limitation, as many investigators do not have access to the compute and storage infrastructure to host datasets of large scale. Instead, many groups have started to use cloud services. There is, however, no cloud-based system that allows for a single point of storage. This creates the wasteful and unworkable situation where researchers must store multiple copies of the same dataset on the same cloud, as there is currently no mechanism to allow the researchers to access a common copy. Solution: 'Bring researchers to the data' - Creation of a pathway for institutions to achieve Trusted Partner status, especially with regards to the utilization of cloud services A number of platforms have emerged to address this challenge and host data on public clouds. However, the community is blocked from using them as there is, in general, no pathway to becoming a trusted partner to distribute data; this is an unnecessary regulatory obstacle that can and should be readily remedied. As a concrete example of how wasteful the current approach is, the NCI has invested significant resources into funding the creation of three "Cloud Pilots," which host TCGA data via a cloud-based platforms (Broad has received one of these awards). Each has Authority To Operate (ATO) as a FISMA Moderate environment, and each is integrated with dbGaP procedures such that, when the DAC grants access to a researcher to utilize TCGA data, it is mirrored in the access control of the cloud pilots. We and others have repeatedly asked for Trusted Partner status to host not only TCGA data through this platform, but also additional data sets that researchers in our community have already placed onto the cloud. Doing this would require no additional funding from the NIH, and it would resolve the current predicament of researchers who must store multiple copies of the same data set on the same cloud, as there is currently no mechanism to allow the researchers to access a common copy. We have been unable to advance this conversation.
- **Benefits and risks associated with the availability of genomic study summary statistics**
- **NIH is interested in public comments on benefits and risks of such reference use of dbGaP data for research participants, patients, and the scientific community**

4. General Comments. NIH welcomes general comments on any other topics with regard to dbGaP data submission, access, and management

We thank the NIH for this thoughtful Request for Information (RFI) on dbGaP Policies and Procedures. We commend dbGaP for furthering the cause of data sharing and appreciate its efforts to expedite the review process in the face of an exponentially growing volume of data submissions and data requests. As one of the largest U.S. genome centers, Broad Institute has had extensive experience with all aspects of dbGaP, ranging from data deposition to data discovery and retrieval. Moreover, members of our community have been involved in crafting the frameworks for data access and data use for initiatives such as such as the All of Us Research Program and the Global Alliance for Genomics and Health, and we have played a key role in data coordination for large-scale projects such as TCGA, 1000 Genomes, ExAC and GTEx. We are thus well-positioned to comment on the current state and future role of dbGaP. In crafting this response, we have attempted to draw upon these experiences and summarize the perspectives of the Institution at large. Our fundamental view is that the community desperately needs improved mechanisms to support data sharing. Given the social investment that has gone into generating the data currently stored in dbGaP, the paucity of investment into technologies for making it usable seems misguided. The ethical discussion surrounding data sharing needs to be reframed to recognize that there is an ethical obligation to make these data more easily accessible to responsible researchers. In this response, we offered a solution for each problem we have identified in numbers 1 through 3. We welcome the opportunity to work with NIH and the community to help implement these solutions. In particular, two general areas stand out to us as urgently in need of change:
1. Genomics researchers are beginning to make greater use of cloud services, yet dbGaP is not creating Trusted Partner or other mechanisms to allow for groups to host data on the cloud (with the exception of a few thoughtful programs such as the NCI Cloud Pilots). This creates the wasteful and unworkable situation where researchers must store multiple copies of the same dataset on the same cloud, as there is currently no mechanism to allow the researchers to access a common copy. 2. Enormous amounts of time are spent by Data Access Committees (DACs) policing whether a given data access request is consistent with a given data use restriction. This process is extraordinarily frustrating as there is a high degree of variability on granting access depending on the DAC. It is our view that data use restrictions can be structured as ontologies, so that this consistency checking can be automated by computers.
Were this done, researchers would be able to search for datasets that are consistent with their research purposes. The work of DACs would be substantially reduced, allowing for focus on of a minority of harder requests. We provided a solution to this problem.

Additional Document

Broad Institute Response to RFI on dbGAP

Introductory Remarks - will have to go in “Other”

We thank the NIH for this thoughtful Request for Information (RFI) on dbGaP Policies and Procedures. We commend dbGaP for furthering the cause of data sharing and appreciate its efforts to expedite the review process in the face of an exponentially growing volume of data submissions and data requests.

As one of the largest U.S. genome centers, Broad Institute has had extensive experience with all aspects of dbGaP, ranging from data deposition to data discovery and retrieval. Moreover, members of our community have been involved in crafting the frameworks for data access and data use for initiatives such as such as the *All of Us* Research Program and the Global Alliance for Genomics and Health, and we have played a key role in data coordination for large-scale projects such as TCGA, 1000 Genomes, ExAC and GTEx. We are thus well-positioned to comment on the current state and future role of dbGaP. *In crafting this response, we have attempted to draw upon these experiences and summarize the perspectives of the Institution at large*.

Our fundamental view is that the community desperately needs improved mechanisms to support data sharing. Given the social investment that has gone into generating the data currently stored in dbGaP, the paucity of investment into technologies for making it usable seems misguided. The ethical discussion surrounding data sharing needs to be reframed to recognize that there is an **ethical obligation** to make these data more easily accessible to responsible researchers.

In this response, we offered a **solution** for each **problem** we have identified in numbers 1 through 3. We welcome the opportunity to work with NIH and the community to help implement these solutions.

In particular, two general areas stand out to us as urgently in need of change:

1. Genomics researchers are beginning to make greater use of cloud services, yet dbGaP is not creating Trusted Partner or other mechanisms to allow for groups to host data on the cloud (with the exception of a few thoughtful programs such as the NCI Cloud Pilots). This creates the wasteful and unworkable situation where researchers must store multiple copies of the same dataset on the same cloud, as there is currently no mechanism to allow the researchers to access a common copy.
2. Enormous amounts of time are spent by Data Access Committees (DACs) policing whether a given data access request is consistent with a given data use restriction. This process is extraordinarily frustrating as there is a high degree of variability on granting access depending on the DAC. It is our view that datause restrictions can be structured as ontologies, so that this consistency checking can be automated by computers. Were this done, researchers would be able to search for datasets that are consistent with their research purposes. The work of DACs would be substantially reduced, allowing for focus

on of a minority of harder requests. We provided a solution to this problem.

1. Study Registration and Submission

Problem: The study registration and submission process is overly complex, manual, lacks a systematic manner in which data is curated, and does not accurately capture critical phenotypic information.

Solution:

1. Automate sample registration process. Much of the sample registration process is manual and overly time consuming. Currently, this registration requires communicating directly with a specific dbGaP officer, which is not scalable.

2. Develop simple tools for the deposition of phenotypic data. Phenotypes are not structured on deposition, greatly complicating meaningful secondary use. NIH must either develop or encourage others to develop tools for structuring phenotypes. These structured phenotypes must be part of a comprehensive ontology system to enable flexible queries and searching.

3. Require deposition of all phenotypic data that is published. Again, meaningful secondary use requires good phenotypic data. If investigators are funded by NIH and have the phenotypic data and the tools are provided for easy deposition (#2 above), NIH should require the deposition of the data (at minimal, the phenotypic data that was already included in a publication).

4. Data Submission and Curation. There exist several groups with expertise in large-scale data engineering, as a result of efforts such as TCGA, GTEx, ExAC and TOPMed. This expertise should be leveraged by creating a small number of centers (but more than one), that are charged with
    a. Processing of genomic data with various best practices pipelines. This mandate should include reprocessing data with new builds of the genome, or when there is an algorithmic advance in variant calling significant enough to justify it.
    b. Mapping phenotypes to ontologies.
    c. Structuring data use restrictions so that they are mapped to a standard ontology (see Section 2, below).

5. Computing Infrastructure: The curated datasets should be placed onto public clouds for use by the community (more than one is crucial to maintain competition), in addition to NIH-owned servers from which it can be downloaded. It is crucial that these datasets be directly accessible by users, and that users not be required to go through any given platform to access a given dataset.

6. Access Control: dbGaP should continue to adjudicate access to publicly funded datasets. Under this model, dbGaP would continue to be an identity provider that adjudicates who is a “trusted researcher,” and researchers would continue to apply to this body to receive permissions to access various datasets. The data should be stored on one or more cloud infrastructures; as access is granted, this would be mirrored across infrastructures. dbGaP should make whitelists of which researchers are allowed to access the various applications publicly available via APIs so that third-party applications could respect these permissions.

   In summary, the above model would facilitate access:
   a. By ceding data curation to groups with expertise in it, this work would no longer fall squarely on the shoulders of dbGaP.
   b. By making data use restrictions machine-readable, the work of DACs would be greatly simplified.
   c. By utilizing clouds, there would be less need for downloading data to new environments, which significantly taxes the current system.

2. Data Access Request and Review

Problem: The system for granting data access is entirely manual. This process is both: (a) highly variable based on the Data Access Committee; (b) simply not scalable. Additionally, the annual renewal and reporting process is extraordinarily burdensome for investigators to complete and NIH staff to review and provides little benefit.

Solutions:

I. Standardize Data Use Letters (DULs) and Data Access Requests (DARs).

The Broad has developed a user friendly tool and has piloted this experiment. We inspected a collection of nearly 132 DULs at the Broad. From that exercise, we determined that 95% could be structured into an ontology that contained the following 5 main categories (see Figure below): i) disease-specific restrictions, ii) commercial restrictions, iii) restrictions to special populations, iv) restrictions on research use, v) General Research Use.

If DULs were structured to follow a standard ontology, researchers would be able to search for datasets consistent with their research purpose. It would also greatly facilitate the work of the DAC, as the effort of checking whether a DAR is consistent with a DUL would be automatable and thus scalable. Towards this end, our Institute has developed an open-source software package (“DUOS”) for structuring DULs and filtering them by research purpose. This software package also contains interfaces for DACs to facilitate review of data access requests.

We would happily donate this software to dbGaP for use by its DACs.

II. Phenotype-based search

Just as the previous example noted the need to data-use-enabled search, there is a significant need for phenotype-based search. Researchers would greatly benefit from the ability to search for all samples with a given phenotype in constructing cohorts. This search should be ontology-aware (e.g., search for “cancer” includes “angiosarcoma” etc’).

III. Extending and Renewing Data Access Requests
Currently, if a researcher has approval to study one dataset and later wants to utilize a newly deposited dataset for the same research, this entails an entirely new application. There should be a mechanism by which researchers can extend an already-approved application to include additional datasets as they emerge.

Finally, the annual renewal process is unnecessarily burdensome for both Investigators and NIH staff. We propose that all investigators agree to a code of conduct rather than annual renewal and that an audit function by the “trusted partner” replace annual renewals.

IV. Better coordination of DARs with Cohorts
Many of the traditional large-scale cohorts are governed by an added layer of access control--either a local IRB or a DAC that is affiliated with the cohort. This can greatly complicate and prolong the process of accessing data, and we have experienced inconsistency between cohorts with regard to protocols and policies.

We propose that this additional layer be streamlined where one representative from each cohort is designated to make decisions in a time-limited fashion. And, if there are additional limitation on data use, the group should make them entirely transparent so that researchers will only ask for the data if they can meet any of the additional layers of review.

3. Management and Use of dbGaP data

Problem 1: ‘Bring data to the researchers’- researchers are forced to make copies of the data. dbGaP presumes that researchers will download data to their own infrastructure. **Accessibility** is a significant limitation, as many investigators do not have access to the compute and storage infrastructure to host datasets of large scale. Instead, many groups have started to use cloud services. There is, however, no cloud-based system that allows for a single point of storage. This creates the wasteful and unworkable situation where researchers must store multiple copies of the same dataset on the same cloud, as there is currently no mechanism to allow the researchers to access a common copy.

Solution: ‘Bring researchers to the data’ - Creation of a pathway for institutions to achieve Trusted Partner status, especially with regards to the utilization of cloud services

A number of platforms have emerged to address this challenge and host data on public clouds. However, the community is blocked from using them as there is, in general, no pathway to becoming a trusted partner to distribute data; this is an unnecessary regulatory obstacle that can and should be readily remedied.

As a concrete example of how wasteful the current approach is, the NCI has invested significant resources into funding the creation of three “Cloud Pilots,” which host TCGA data via a cloud-based platforms (Broad has received one of these awards). Each has Authority To Operate (ATO) as a FISMA Moderate environment, and each is integrated with dbGaP procedures such that, when the DAC grants access to a researcher to utilize TCGA data, it is mirrored in the access control of the cloud pilots.

We and others have repeatedly asked for Trusted Partner status to host not only TCGA data through this platform, but also additional datasets that researchers in our community have already placed onto the cloud. *Doing this would require no additional funding from the NIH, and it would resolve the current predicament of researchers who must store multiple copies of the same dataset on the same cloud, as there is currently no mechanism to allow the researchers to access a common copy.* We have been unable to advance this conversation.

4. Additional Comments

NOTE TO GSG - We will have to put the intro here as there is no other place for it.

Submission Date
04/07/2017
Name
Brandi Davis-Dusenbery
Primary Purpose of dbGaP Use
Both Study Registration / Data Submission and Data Access / Download
What is your level of experience with dbGaP?
Experienced (used many times over the course of several years)
Role/Other Role
Institutional Official
Type of Organization/Other Type
Biomedical software developer
Name of Organization
Seven Bridges Genomics

Information Requested

1. dbGaP Study Registration and Data Submission
Seven Bridges has provided assistance to multiple research groups for whom reprocessing restricted data and publishing the results posed particular challenges, particularly for those developing novel algorithms. Currently, the determination of whether newly processed data is restricted is performed on a largely ad hoc basis. Providing more formal guidelines and, where possible, explicit methodologies for testing the data produced would be extremely helpful. For example, new methods for identifying somatic variants almost invariably uncover germline variants as well. As this case is relatively common, test datasets with known results could be provided to assist researchers and the Data Access Committee in evaluating to what extent private information could be released and the associated risk levels. This would allow researchers to better understand and plan the timelines for publication and research. Once data is processed, it can be challenging to determine best-practices for sharing the results (either publicly or in a restricted manner). While high-level flow charts exist that cover the process of creating a new dbGaP project, we were unable identify a step-by-step method for submitting reprocessed data as a dbGaP project for publication. Furthermore, we were contacted by several research groups using for advice on the same issue. If such documentation exists, it would be helpful to more strongly highlight it. Even better would be to provide best-practice publication guidelines for newly created public and restricted data sets. These best-practices could include a single, recommended point for public data to be gathered for future comparisons or meta-analyses.

2. dbGaP Data Access Request (DAR) and Review
As an NCI Cancer Genomics Cloud pilot provider (available at www.cancergenomicscloud.org), Seven Bridges has experience submitting data access requests for multiple dbGaP regulated studies. Additionally, as a NIH Trusted Partner, Seven Bridges is frequently the first point of contact for researchers hoping to access controlled access data (ie TCGA, TARGET, CGCI). In the approximately 2 years since launching the CGC we have received 3-10 inquiries per month about how researchers should navigate the dbGaP approval processes. To assist answering these questions we have developed a comprehensive tutorial and flow chart available at http://docs.cancergenomicscloud.org/docs/tcga-data-access. This chart captures the 10+ steps needed for researchers to gain approval for controlled data. We frequently interact with researchers who are simply unable to move their science forward due to delays in data approval or insufficient understanding of the process. While streamlining policies and procedures represents an important step, we also believe that incorporation of modern user experience and interaction principles would enhance the user interfaces of the dbGaP system. Importantly, providing additional (including visual as well as programmatic) methods for researchers to discover what data is available will be critical to facilitate data use. We also note that we've had the opportunity to interact with members of multiple data access committees and have been deeply impressed by the dedication and rigour that these individuals bring to ensure data are used properly and efficiently.

3. Policies for the Management and Use of dbGaP Data

- **Alternate controlled-access models**
- **Benefits and risks associated with the availability of genomic study summary statistics**
- **NIH is interested in public comments on benefits and risks of such reference use of dbGaP data for research participants, patients, and the scientific community**
  Researchers seeking to compare their data to controlled-access public datasets are often interested in clinically

relevant subsets the data. For this reason, it would be helpful to provide the ability for datasets to include an extensible metadata schema. As a concrete example, several research groups have independently performed HLA typing of TCGA datasets. These groups could each provide a publication-associated metadata extension for the relevant TCGA datasets that other researchers could use in their own analyses. By making the metadata extensible, as opposed to requiring replacement of existing metadata, researchers could update file metadata with new methods, compare the results of multiple existing methods, and perform meta-analyses and more complex analyses of multi-dimensional data. Ideally, the extended metadata would be curated and made searchable to help researchers increase their statistical power by producing in silico cohorts. The utility of analyzing clinical data alongside public data relies on the ability to use the same analytical methods to generate processed results. Ideally, processed data in dbGaP projects would provide the pipelines and references used to produce the processed data, as well as suggest methods for data harmonization. This would reduce the burden on clinical researchers seeking to compare their results to public datasets.

4. General Comments. NIH welcomes general comments on any other topics with regard to dbGaP data submission, access, and management

Improving the ability of research groups to search across multiple datasets will become increasingly important. Metadata search functionality will need to be accessible to users with varying degrees of comfort with the command-line and database query languages. Seven Bridges and others, such as the GDC, have provided initial forays into assisting users with querying datasets using both visual interfaces and API. We have observed that researchers can be confused by the complex metadata schema used to organize large datasets. While such complexity is challenging to avoid in large clinical datasets, it may be useful to begin developing core, structured components of metadata schema for use in public datasets from biological samples. While the complexity of metadata schema would not be reduced, the schema could at least become familiar and more easily queried across diverse datasets.

Submission Date
04/07/2017
Name
Mary Majumder for Project Team
Primary Purpose of dbGaP Use
What is your level of experience with dbGaP?
Role/Other Role
Bioethicist
Type of Organization/Other Type
University
Name of Organization
Center for Medical Ethics and Health Policy, Baylor College of Medicine Other Type of Organization

Information Requested

1. dbGaP Study Registration and Data Submission

2. dbGaP Data Access Request (DAR) and Review
Per current DbGaP procedures, Data Access Committees are made up exclusively of government employees. This government-official only policy is understandable given history, but it is also a serious long-term liability, indeed a flaw, as it excludes some of the most credible advocates, patients, survivors, previvors and other constituencies who should be part of the governance structure. It also means that changes within the government can change policy over management of data in ways that the public would not support, and these data initiatives would be better trusted and more stable if nongovernment personnel were among the gate-keepers and overseers. This does raise issues of conflict-of-interest and criteria for selecting representatives, but those are real problems of real politics that cannot be avoided, and indeed governance structures may be a way of confronting such politics early and effectively. For more detail, see the attached document with our full comments.

3. Policies for the Management and Use of dbGaP Data

- **Alternate controlled-access models**
- **Benefits and risks associated with the availability of genomic study summary statistics**
  Presumably, most of the data within the category of genomic study summary statistics would not create significant re-identification risk if made public, does not focus on a community or set of communities with special research-related concerns, and is unlikely to be used in studies on sensitive research topics (see the attached document with our full comments for more information). Further, the potential benefits of open access to such data—which could still be subject to stated conditions such as no attempts at re-identification—appear to be considerable. Nevertheless, some of our interviewees forcefully articulated the view that "expert" risk-benefit analysis alone is insufficient to resolve questions that involve values and trade-offs impacting participants. Accordingly, they would urge the research community, including NIH, to embrace the opportunity to build a component of systematic engagement of participants into deliberations about these important policy questions (see the attached document with our full comments for more information concerning the importance of engagement of participants, as well as communities and the general public).
- **NIH is interested in public comments on benefits and risks of such reference use of dbGaP data for research participants, patients, and the scientific community**

4. General Comments. NIH welcomes general comments on any other topics with regard to dbGaP data submission, access, and management

Additional Document

Re Notice Number NOT-OD-17-044

We appreciate the opportunity to comment on topics related to the NIH National Center for Biotechnology Information database of Genotypes and Phenotypes (dbGaP) in response to the Request for Information released February 21, 2017 (RFI), especially "Policies for the Management and Use of dbGaP Data." In formulating the following comments, we draw on preliminary findings from research we are conducting in connection with the Building the Medical Information Commons: Participant Engagement and Policy project (R01 HG008918).

Importance of data sharing for multiple purposes

Research validates the belief that broad data sharing fuels scientific productivity. Human genes initially sequenced by Celera Corporation and thereafter maintained as proprietary resources were cited by 20%-30% fewer research papers and led to fewer diagnostic tests for those genes than the genes first mapped by the Human Genome Project and made public under the Bermuda Principles.[1] Controlled access, like maintenance of data in proprietary databases, has the potential to diminish scientific productivity. Open-access data sets of human genomic variation (e.g., HapMap, 1000 Genomes, HapMap, ExAC, GenBank, RefSeq, ClinVar, BRCA Exchange, CFTR2) are used by many more researchers than related data sets within dbGaP.[2] Concerning genomic study summary statistics in particular, the ExAC open access genetic variation database has been singled out as a resource that has proven immensely valuable for both research and clinical studies.[3]

Importance of participant/community/public engagement

The growing recognition of the value of broad data sharing has been accompanied by calls for more engagement of participants, communities, and the public. In the case of participants, one factor is widespread attention to the risk of re-identification of individual participants in the era of "big data."[4] Indeed, the decision to manage data within dbGaP according to a controlled access model was a response to work establishing the technical feasibility of re-identification of at least some individual research participants.[5] In addition, some commentators are making the case for more attention to *inference risk*, which has been defined as "the potential for others to

[1] Williams H. Intellectual property rights and innovation: evidence from the human genome. *Journal of Political Economy* 2013;121(1):1-27.

[2] Rodriguez LL, Brooks LD, Greenberg JH, Green ED. The complexities of genomic identifiability. *Science* 2013;339:275-276.

[3] Bahcall OG. ExAC boosts clinical variant interpretation in rare diseases. *Nature Reviews Genetics* 2016;17: 584.

[4] For example, a recent letter reporting the findings of the National Committee on Vital and Health Statistics (NCVHS) regarding de-identification under HIPAA cites expert testimony suggesting that "the goals of preserving the individual's right to privacy while fully using digital information to improve health and outcomes may be on a 'collision course.'" The NCVHS also notes that genomic data "present unique de-identification challenges." William W. Stead, Chair, NCVHS, Letter to Hon. Thomas E. Price, Secretary, Department of Health and Human Services, Re: Recommendations on De-identification of Protected Health Information under HIPAA, dated February 23, 2017, available at: https://www.ncvhs.hhs.gov/wp-content/uploads/2013/12/2017-Ltr-Privacy-DeIdentification-Feb-23-Final-w-sig.pdf.

[5] Paltoo DN et al. Data use under the NIH GWAS Data Sharing Policy and future directions. *Nature Genetics* 2014;46:934-938.

learn about individuals from their inclusion in a dataset, or from their membership in, or association, or perceived association, with the group studied, even if the individual's actual data was not included in the data set."[6]

In the face of these challenges, one option is to focus on the informed consent process, and, in the case of existing data sets, recontact and reconsent individual participants for broad sharing in an environment in which privacy and confidentiality cannot be guaranteed. Another option, also supported by a strong ethical justification, is to put questions of benefit and risk to a multidisciplinary committee that includes patient advocates or participant representatives.[7] This option aligns with a more general trend over the last several decades: increasing attention to governance structures and the processes by which decisions are made about aspects of research including future uses of resources generated through research, and advocacy for more participant involvement in such structures and processes.[8] This trend is not merely a response to new kinds or levels of risk; it grows out of continuing reflection on the implications of the principle of respect in the context of research.

Calls for greater involvement of communities are likewise the outgrowth of multiple developments, including a series of scandals involving research in communities that have a history of unjust treatment or are vulnerable to exploitation (or both), the potential for harm that transcends individual participants (as suggested by the concept of inference risk), new understandings of respect that extend to communities as well as individuals, and theories of deliberative democracy. While some communities are readily identified, relatively cohesive, and have internal processes for designating some individuals as representatives authorized to speak on their behalf, other communities may have fuzzier boundaries and looser organization. In the case of American Indian and Alaska Native communities, the National Congress of American Indians has made available extensive resources to support those communities in considering various aspects of genetic research, including data sharing, and a primer for outside researchers and others on how to engage with these communities in an informed and respectful manner.[9] Models also exist for engaging other kinds of communities, providing guidance on associated challenges and how to manage them.[10]

Finally, another option is engagement with the public through democratic practices. The case for public engagement rests on theories of deliberative democracy and the positive results of experiments with deliberative methods involving members of the general public.[11]

[6] NCVHS Letter, supra note 4.

[7] Peppercorn J. et al. Ethical aspects of participation in the Database of Genotypes and Phenotypes of the National Center for Biotechnology Information: the Cancer and Leukemia Group B experience. *Cancer* 2012;118:5060-5068.

[8] E.g., Caulfield T, Einsiedel E, Merz JF, Nicol D. Trust, patents and public perceptions: The governance of controversial biotechnology research. *Nature Biotechnology* 2006;24(11):1352-4; Allyse MA, McCormick JB, Sharp RR. Prudentia populo: involving the community in biobank governance. *American Journal of Bioethics* 2015;15:1-3.

[9] National Congress of American Indians. American Indian & Alaska Native Genetics Resource Center. Available from: http://genetics.ncai.org.

[10] E.g., Joosten YA et al. Community engagement studios: a structured approach to obtaining meaningful input from stakeholders to inform research. *Academic Medicine* 2015;90:1646-1650; Allyse, McCormick & Sharp, supra note 8.

[11] Carman KL et al. Public deliberation to elicit input on health topics: Findings from a literature review. Rockville, MD: Agency for Healthcare Research and Quality, 2013. Available from:

Understanding the views of members of the public is valuable in planning programs that require public support. This is especially important when policy decisions entail value-sensitive questions or involve controversy, which many questions surrounding genomics do. Public engagement should lead to policies that are more acceptable to the public than those developed by technical experts alone.[12]

As part of our NHGRI-funded “Building the Medical Information Commons” project, which focuses on ethical and policy issues related to the development of large-scale data resources for research, clinical, and other uses, we interviewed 22 expert stakeholders from several sectors (e.g., academia, technology industry, government). *All* respondents expressed support for greater participant engagement. Some noted that including participants in governance may lead to less rather than more restrictive data-sharing policies: “[T]he people whose data it is can be outstanding advocates for how that information can and should be more broadly accessed and shared, as well as of course expressing when they’re having concerns about the sharing of that data,” said one of the respondents. Others acknowledged challenges related to recruitment of individuals to represent or reflect large and potentially diverse populations but have concluded that these challenges are not insurmountable, as illustrated by this quote:

> “[P]articipant leadership is important in the governance structure, and… it seems to me that’s the way to realize the autonomy principle, more in a collective way than in an individual way. Of course that raises concerns about who would represent the participants... We have mechanisms of thinking about that in the context of democratic governance more generally. That’s how I would think about participant leadership in the governance structure.”

Finally, respondents emphasized that meaningful engagement requires meaningful investment of time, energy, and resources:

> “[I]f we want real people and we’re not looking for people that are already a hundred percent bought into science… [it’s important] that we give them an opportunity through orientation training and ongoing support to really have a voice and to actively participate. Not everybody does that. They might throw somebody onto an advisory board, but not actually support them.”

Respondents also reflected on the importance of enlisting participants and communities in identifying research priorities and in decision making about the use of their data in categories of research they might consider outside the scope of “biomedical research” and/or of special concern (e.g., risk of affecting perceptions of an entire tribe or minority group). A number of articles contain lists of sensitive research topics and/or categories of sensitive data.[13] The

http://www.effectivehealthcare.ahrq.gov/ehc/assets/File/Deliberation-Public-Lit-Review-130213.pdf; Carman KL et al. Community forum deliberative methods demonstration: Evaluating effectiveness and eliciting public views on use of evidence. Rockville, MD: Agency for Healthcare Research and Quality, 2014; Garrett SB, Dohan D, Koenig BA. Linking broad consent to biobank governance: support from a deliberative public engagement in California. Am J Bioeth. 2015 Sep; 15(9):56-7. PMID: 26305757; Garrett SB, Koenig BA, Brown A, Hult JR, Boyd EA, Dry S, Dohan D. EngageUC: developing an efficient and ethical approach to biobanking research at the University of California. Clin Transl Sci. 2015 Jan 10. PMID: 25581047.

[12] See, e.g., Tomlinson T, De Vries R, Ryan K, Kim HM, Lehpamer N, Kim SY. Moral concerns and the willingness to donate to a research biobank. *JAMA* 2015;313(4):417-9; Caulfield, Einsiedel, Merz & Nicol, supra note 8.

[13] Hayden EC. Taboo genetics. *Nature* 2013;502:26-28; Lo B, Barnes M. Federal research regulations for the 21st century. *New England Journal of Medicine* 2016;374:1205-1207; Dyke SOM, Dove ES, Knoppers BM. Sharing health-related data: a privacy test? *Genomic Medicine* 1, Article number: 16024 (2016) doi:10.1038/npjgenmed.2016.24; DeVries R et al., The moral concerns of biobank donors: the effect of non-welfare

typology of categories of sensitive data, which includes “genetic” data as one category of sensitive data, is accompanied by a test for determining which types of information within sensitive data categories should fall within a zone of heightened protection.[14] While valuable, this work would be advanced by more extensive engagement with the individuals and groups affected. One respondent noted that “we have a considerable lack of data about what people who have been comfortable to put their data into data repositories actually think is going to happen with their data.” This respondent commented on recent work on the genetics of intelligence as an example of research that might violate participant expectations and trust in the absence of explicit informed consent.[15]

“Low-risk” data

Presumably, most of the data within the category of genomic study summary statistics would not create significant re-identification risk if made public, does not focus on a community or set of communities with special research-related concerns, and is unlikely to be used in studies on sensitive research topics. Further, the potential benefits of open access to such data—which could still be subject to stated conditions such as no attempts at re-identification—appear to be considerable. Nevertheless, some of our interviewees forcefully articulated the view that “expert” risk-benefit analysis alone is insufficient to resolve questions that involve values and trade-offs impacting participants. Accordingly, they would urge the research community, including NIH, to embrace the opportunity to build a component of systematic engagement of participants into deliberations about these important policy questions.

Additional observations

As a final set of observations, interviews and our other explorations of building a medical information commons have identified several potential flash-points for future problems. One is international data sharing. Given evidence of potential public resistance to sharing of genomic and other health-related data across national borders, we believe it is essential that policymakers and leaders in the scientific community make the case for international data sharing *to the public.*[16] Another is distrust of government. dbGaP, GenBank, ClinVar, and many other databases are U.S.-government-owned and operated. In the long run, this may prove difficult to sustain in the face of international distrust of U.S. institutions, or any one government’s control

interests on willingness to donate. *Life Sciences, Society and Policy* 2016;12 doi10.1186/s40504-016-0036-4; Tomlinson, De Vries, Ryan, Kim, Lehpamer & Kim, supra note 12.

[14] Dyke, Dove & Knoppers, supra note 13.

[15] Recently published work on the genetics of educational attainment and cognition includes: Kong A et al. Selection against variants in the genome associated with educational attainment. *Proceedings of the National Academy of Sciences* 2017; 114: E727–E732; Trampush JW et al. GWAS meta-analysis reveals novel loci and genetic correlates for general cognitive function: a report from the COGENT consortium. *Molecular Psychiatry* 2017;22:336-345; Okbay A et al. Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* 2016;533:539-542; Davies G et al. Genome-wide association study of cognitive functions and educational attainment in UK Biobank. *Molecular Psychiatry* 2016;21:758-767. For insight into the controversial nature of this research, see Parens E, Appelbaum PS. An introduction to thinking about trustworthy research into the genetics of intelligence. *Hastings Center Report* 2015;45(5):S2-S and other articles contained in this special supplement.

[16] Majumder MA, Cook-Deegan R, McGuire AL. Beyond our borders? Public resistance to global genomic data sharing. *PLoS Biology*. 14(11): e2000206. doi: 10.1371/journal.pbio.2000206.

(not just the United States). Moreover, distrust is intensified when the gate-keepers to the database consist entirely of government employees. Per current DbGaP procedures, Data Access Committees are made up exclusively of government employees. This government-official only policy is understandable given history, but it is also a serious long-term liability, indeed a flaw, as it excludes some of the most credible advocates, patients, survivors, previvors and other constituencies who should be part of the governance structure. It also means that changes within the government can change policy over management of data in ways that the public would not support, and these data initiatives would be better trusted and more stable if nongovernment personnel were among the gate-keepers and overseers. This does raise issues of conflict-of-interest and criteria for selecting representatives, but those are real problems of real politics that cannot be avoided, and indeed governance structures may be a way of confronting such politics early and effectively. Finally, rules that are overly focused on compliance with the consent of the original data donor do not allow data-gathering organizations to take a strong position in establishing oversight. This presents a particular problem for organizations like tribal governments, but applies to other data collectors as well.

Thank you for your consideration of our comments,

Juli Bollinger, MS, Robert Cook-Deegan, MD, Patricia Deverka, MD, MS, MBE, Barbara Koenig, PhD, Mary A. Majumder, JD, PhD, Amy L. McGuire, JD, PhD, Angela Villanueva, MPH, CPH

Submission Date
04/07/2017
Name
Sarah Nelson
Primary Purpose of dbGaP Use
Study Registration / Data Submission
What is your level of experience with dbGaP?
Experienced (used many times over the course of several years)
Role/Other Role
Scientific Researcher
Type of Organization/Other Type
University
Name of Organization
University of Washington

Information Requested

1. dbGaP Study Registration and Data Submission

2. dbGaP Data Access Request (DAR) and Review 3. Policies for the Management and Use of dbGaP Data

- **Alternate controlled-access models**
- **Benefits and risks associated with the availability of genomic study summary statistics**
  I agree with the Summary from the NHGRI Workshop on Sharing Aggregate Genomic Data that placing genomic summary statistics behind controlled access creates an "unnecessary…and undesirable" byproduct of giving more users access to individual-level genotype data than actually want or need it. However, to make all genomic summary statistics publicly available might be an over correction. Instead, I think study investigators should play a more active role in determining whether or not summary statistics *for their participants* should be in controlled vs. uncontrolled access portions of dbGaP. Investigators arguably have the closest relationship to participants and are thus better stewards (vs. a blanket dbGaP policy) about what would and wouldn't be acceptable to their participants. I also noticed much of the Workshop Summary focused on risks to participants. An alternate framing is to focus on respecting what participants understood would be happening with their data. Which again circles back to the point that perhaps study investigators should be deciding whether summary stats can be publicly available for their given study. Then determinations of "vulnerable populations" and "sensitive phenotypes" can be adjudicated by the study team, not by dbGaP as whole.
- **NIH is interested in public comments on benefits and risks of such reference use of dbGaP data for research participants, patients, and the scientific community**

4. General Comments. NIH welcomes general comments on any other topics with regard to dbGaP data submission, access, and management

Submission Date
04/07/2017
Name
Mark Daly
Primary Purpose of dbGaP Use
Both Study Registration / Data Submission and Data Access / Download
What is your level of experience with dbGaP?
Experienced (used many times over the course of several years)
Role/Other Role
Scientific Researcher
Type of Organization/Other Type
Nonprofit Research Organization
Name of Organization
Broad Institute

Information Requested

1. dbGaP Study Registration and Data Submission

2. dbGaP Data Access Request (DAR) and Review

3. Policies for the Management and Use of dbGaP Data

- **Alternate controlled-access models**
- **Benefits and risks associated with the availability of genomic study summary statistics**
  Problem: There is minimal risk of harm to research participants from openly sharing genetic study summary statistics. Conversely, the benefit to openly sharing this information is enormous. When any ethics committee reviews a research protocol, it weighs the risk and the benefits, and here there is no doubt open access sharing offers benefits far exceeding risks. In specific: 1) Risk of re-identification: The release of summary statistics from a genetic study can only allow a researcher to determine if a specific individual was a 'case' participant of the study (thereby revealing the 'case' status for that disease) if and only if the researcher has access to that individual's genome already. This is minimal risk: what parties, without access to your medical information and diagnostic status, would nonetheless acquire a tissue specimen and sequence it at great cost? Further, why would they then conduct complex analytic decomposition to discover the unlikely fact that you had participated in a published genetic study as a case? 2) Likelihood that re- identification would be pursued: There is little incentive for anyone to re-identify specimens other than to demonstrate that they can do so. Accordingly, to date, there are no instances of re-identification of data or specimens for illicit motives [CITI Training – CITI Genetics Workgroup]. Moreover, researchers obtaining access to some de-identified datasets (e.g., dbGAP) are required to attest and guarantee they will not attempt to re-identify data. 3) Matching 'Data Use Restrictions' to a research purpose as a prerequisite to granting access for summary statistics minimizes the overall benefit to research participants: We argue that it is highly unlikely, considering the proven altruism of research participants and their desire to have researchers study their conditions, that they would not want their data used in aggregate analysis. For example, it is difficult to imagine a scenario in which a research participant who consented for a cancer study would be opposed to that data being used on a summary level to build a more accurate human genome that might allow advances in research beyond cancer. 4) Sharing summary statistics has tremendous benefit to advancing disease research: Specifically, the broadest possible sharing of summary statistics promotes the development of statistical methods, the progress of biomedical research into these specific diseases, and advances fundamental discoveries in genetics and biology. 5) Sharing summary statistics benefits individuals with rare diseases: It allows comparison of their genomic profiles to the general population and enables researchers and clinicians to narrow the list of potential disease-causing variants (for example, as in the use of ExAC and gnomAD). Solution: (a) Open access - As outlined above, based on the risk and benefit, there is no ethically justifiable reason not to allow open access. (b) Registered access - in the unfortunate case there must be registered access we recommend a model where the researcher need not specify their research purpose but only attest to a code of conduct. Additionally, one entity that has consistent methodology for approving requests is critical (unlike the current dbGaP Data Access Committees).
- **NIH is interested in public comments on benefits and risks of such reference use of dbGaP data for research**

participants, patients, and the scientific community

4. General Comments. NIH welcomes general comments on any other topics with regard to dbGaP data submission, access, and management

Additional Comment

Benefits and risks associated with the availability of genomic study summary statistics

Based on Mark Daly's ,
https://docs.google.com/document/d/1W_U7LiHy6VB6tCqfsvTaa9XzsM0__0G5NwldZOBX1AA/edit

Problem:
There is minimal risk of harm to research participants from openly sharing genetic study summary statistics. Conversely, the benefit to openly sharing this information is enormous. When any ethics committee reviews a research protocol, it weighs the risk and the benefits, and here there is no doubt open access sharing offers benefits far exceeding risks.

In specific:

1) Risk of re-identification: The release of summary statistics from a genetic study can only allow a researcher to determine if a specific individual was a 'case' participant of the study (thereby revealing the 'case' status for that disease) if and only if the researcher has access to that individual's genome already. This is minimal risk: what parties, without access to your medical information and diagnostic status, would nonetheless acquire a tissue specimen and sequence it at great cost? Further, why would they then conduct complex analytic decomposition to discover the unlikely fact that you had participated in a published genetic study as a case?
2) Likelihood that re-identification would be pursued: There is little incentive for anyone to re-identify specimens other than to demonstrate that they can do so. Accordingly, to date, there are no instances of re-identification of data or specimens for illicit motives [CITI Training – CITI Genetics Workgroup]. Moreover, researchers obtaining access to some de-identified datasets (e.g., dbGAP) are required to attest they will not attempt to re-identify data.
3) Matching 'Data Use Restrictions' to a research purpose as a prerequisite to granting access for summary statistics minimizes the overall benefit to research participants: We argue that it is highly unlikely, considering the proven altruism of research participants and their desire to have researchers study their conditions, that they would not want their data used in aggregate analysis. For example, it is difficult to imagine a scenario in which a research participant who consented for a cancer study would be opposed to that data being used on a summary level to build a more accurate human genome that might allow advances in research beyond cancer.
4) Sharing summary statistics has tremendous benefit to advancing disease research: Specifically, the broadest possible sharing of summary statistics promotes the development of statistical methods, the progress of biomedical research into these specific diseases, and advances fundamental discoveries in genetics and biology.
5) Sharing summary statistics benefits individuals with rare diseases: It allows comparison of their genomic profiles to the general population and enables researchers and

clinicians to narrow the list of potential disease-causing variants (for example, as in the use of ExAC and gnomAD).

Solution:

(a) Open access - As outlined above, based on the risk and benefit, there is no ethically justifiable reason not to allow open access.

(b) Registered access - in the unfortunate case there must be registered access we recommend a model where the researcher need not specify their research purpose but only attest to a code of conduct. Additionally, one entity that has consistent methodology for approving requests is critical (unlike the current dbGaP Data Access Committees).

Submission Date
04/07/2017
Name
Heidi Rehm
Primary Purpose of dbGaP Use
Study Registration / Data Submission
What is your level of experience with dbGaP?
Experienced (used many times over the course of several years)
Role/Other Role
Scientific Researcher
Type of Organization/Other Type
University
Name of Organization
Brigham & Women's Hospital

Information Requested

1. dbGaP Study Registration and Data Submission
Problem: The study registration and submission process is overly complex, manual, lacks a systematic manner in which data is curated, and does not accurately capture critical phenotypic information. Solution: 1. Automate sample registration process. Much of the sample registration process is manual and overly time consuming. Currently, this registration requires communicating directly with a specific dbGaP officer, which is not scalable. 2. Develop simple tools for the deposition of phenotypic data. Phenotypes are not structured on deposition, greatly complicating meaningful secondary use. NIH must either develop or encourage others to develop tools for structuring phenotypes. These structured phenotypes must be part of a comprehensive ontology system to enable flexible queries and searching. 3. Require deposition of all phenotypic data that is published. Again, meaningful secondary use requires good phenotypic data. If investigators are funded by NIH and have the phenotypic data and the tools are provided for easy deposition (#2 above), NIH should require the deposition of the data (at minimal, the phenotypic data that was already included in a publication). 4. Data Submission and Curation.
There exist several groups with expertise in large-scale data engineering, as a result of efforts such as TCGA, GTEx, ExAC and TOPMed. This expertise should be leveraged by creating a small number of centers (but more than one), that are charged with a. Processing of genomic data with various best practices pipelines. This mandate should include reprocessing data with new builds of the genome, or when there is an algorithmic advance in variant calling significant enough to justify it. b. Mapping phenotypes to ontologies. c. Structuring data use restrictions so that they are mapped to a standard ontology (see Section 2, below). 5. Computing Infrastructure: The curated datasets should be placed onto public clouds for use by the community (more than one is crucial to maintain competition), in addition to NIH-owned servers from which it can be downloaded. It is crucial that these datasets be directly accessible by users, and that users not be required to go through any given platform to access a given dataset. 6. Access Control: dbGaP should continue to adjudicate access to publicly funded datasets. Under this model, dbGaP would continue to be an identity provider that adjudicates who is a “trusted researcher,” and researchers would continue to apply to this body to receive permissions to access various datasets. The data should be stored on one or more cloud infrastructures; as access is granted, this would be mirrored across infrastructures. dbGaP should make whitelists of which researchers are allowed to access the various applications publicly available via APIs so that third-party applications could respect these permissions. In summary, the above model would facilitate access: a. By ceding data curation to groups with expertise in it, this work would no longer fall squarely on the shoulders of dbGaP. b. By making data use restrictions machine-readable, the work of DACs would be greatly simplified. c. By utilizing clouds, there would be less need for downloading data to new environments, which significantly taxes the current system.

2. dbGaP Data Access Request (DAR) and Review
Problem: The system for granting data access is entirely manual. This process is both: (a) highly variable based on the Data Access Committee; (b) simply not scalable. Additionally, the annual renewal and reporting process is extraordinarily burdensome for investigators to complete and NIH staff to review and provides little benefit. Solutions: I. Standardize Data Use Letters (DULs) and Data Access Requests (DARs). The Broad has developed a user friendly tool and has piloted this experiment. We inspected a collection of nearly 132 DULs at the Broad. From that exercise, we determined that 95% could be structured into an ontology that contained the following 5 main categories (see Figure below): i) disease-specific restrictions, ii) commercial restrictions,

iii) restrictions to special populations, iv) restrictions on research use, v) General Research Use. If DULs were structured to follow a standard ontology, researchers would be able to search for datasets consistent with their research purpose. It would also greatly facilitate the work of the DAC, as the effort of checking whether a DAR is consistent with a DUL would be automatable and thus scalable. Towards this end, our Institute has developed an open-source software package ("DUOS") for structuring DULs and filtering them by research purpose. This software package also contains interfaces for DACs to facilitate review of data access requests. We would happily donate this software to dbGaP for use by its DACs. II. Phenotype-based search Just as the previous example noted the need to data-use-enabled search, there is a significant need for phenotype-based search. Researchers would greatly benefit from the ability to search for all samples with a given phenotype in constructing cohorts. This search should be ontology-aware (e.g., search for "cancer" includes "angiosarcoma" etc'). III. Extending and Renewing Data Access Requests Currently, if a researcher has approval to study one dataset and later wants to utilize a newly deposited dataset for the same research, this entails an entirely new application. There should be a mechanism by which researchers can extend an already-approved application to include additional datasets as they emerge. Finally, the annual renewal process is unnecessarily burdensome for both Investigators and NIH staff. We propose that all investigators agree to a code of conduct rather than annual renewal and that an audit function by the "trusted partner" replace annual renewals. IV. Better coordination of DARs with Cohorts Many of the traditional large-scale cohorts are governed by an added layer of access control--either a local IRB or a DAC that is affiliated with the cohort. This can greatly complicate and prolong the process of accessing data, and we have experienced inconsistency between cohorts with regard to protocols and policies. We propose that this additional layer be streamlined where one representative from each cohort is designated to make decisions in a time-limited fashion. And, if there are additional limitation on data use, the group should make them entirely transparent so that researchers will only ask for the data if they can meet any of the additional layers of review.

3. Policies for the Management and Use of dbGaP Data

- **Alternate controlled-access models**
  Problem 1: 'Bring data to the researchers'- researchers are forced to make copies of the data. dbGaP presumes that researchers will download data to their own infrastructure. Accessibility is a significant limitation, as many investigators do not have access to the compute and storage infrastructure to host datasets of large scale. Instead, many groups have started to use cloud services. There is, however, no cloud-based system that allows for a single point of storage. This creates the wasteful and unworkable situation where researchers must store multiple copies of the same dataset on the same cloud, as there is currently no mechanism to allow the researchers to access a common copy. Solution: 'Bring researchers to the data' - Creation of a pathway for institutions to achieve Trusted Partner status, especially with regards to the utilization of cloud services A number of platforms have emerged to address this challenge and host data on public clouds. However, the community is blocked from using them as there is, in general, no pathway to becoming a trusted partner to distribute data; this is an unnecessary regulatory obstacle that can and should be readily remedied. As a concrete example of how wasteful the current approach is, the NCI has invested significant resources into funding the creation of three "Cloud Pilots," which host TCGA data via a cloud-based platforms (Broad has received one of these awards). Each has Authority To Operate (ATO) as a FISMA Moderate environment, and each is integrated with dbGaP procedures such that, when the DAC grants access to a researcher to utilize TCGA data, it is mirrored in the access control of the cloud pilots. We and others have repeatedly asked for Trusted Partner status to host not only TCGA data through this platform, but also additional datasets that researchers in our community have already placed onto the cloud. Doing this would require no additional funding from the NIH, and it would resolve the current predicament of researchers who must store multiple copies of the same dataset on the same cloud, as there is currently no mechanism to allow the researchers to access a common copy. We have been unable to advance this conversation.
- **Benefits and risks associated with the availability of genomic study summary statistics**
  I endorse our findings of the National Human Genome Research Institute Workshop on Sharing Aggregate Genomic Data held May 19-20, 2016 1: Genomic summary statistics provide valuable information regarding which variants contribute to biological function and disease. 2: Public access to genomic summary statistics through central resources maximizes their value by enabling vastly more researchers to use genomic data in biomedical studies. 3: Sharing genomic summary statistics publicly would improve the ease of access to this information while reducing the need for access to individual-level datasets. 4: A number of institutions have approved sharing of summary statistics and these resources are highly utilized in the clinical and research communities. 5: Privacy and confidentiality risks posed by genomic summary statistics are distinct from those posed by individual-level data. 6: The degree of privacy harm that might occur related to inappropriate use of genomic summary statistics depends on what additional information is revealed by determining whether an individual participated in a particular research study. 7: There is greater privacy concern when studying potentially stigmatizing traits or vulnerable populations, because the outcomes of any privacy

breach could cause greater harm. Studies of this nature merit additional considerations and any appropriate protections. 8: There is “institutional risk” in public release of genomic summary statistics related to the potential to damage public trust if expectations related to sharing are not clear. 9: The research community needs to better explain the distinction in genomic summary statistics versus individual-level data with regard to the type of information and the associated risks. 10: Transparency is needed for participants regarding plans to share genomic summary statistics from research studies. 11: Privacy enhancing and security technologies can provide useful protections as part of a risk-mitigation strategy. Such technologies might be appropriate in the context of studies involving stigmatizing traits or vulnerable populations.

- **NIH is interested in public comments on benefits and risks of such reference use of dbGaP data for research participants, patients, and the scientific community**
  Access to large stores of case-level data including genotype and phenotype is critical to inform the ongoing use and interpretation of genomic data in the context of clinical care. Most genetic variants of clinical consequence are extremely rare and therefore access to large genomic studies to increase the chance of identifying these rare cases that may be informative in only a handful of individuals. However, medical access is different than research access in that most uses are small queries for single or a few variants with ongoing access needed as new clinical questions continually arise. Therefore a DAC model requiring review of every indication is an untenable model for the clinical care setting.

4. General Comments. NIH welcomes general comments on any other topics with regard to dbGaP data submission, access, and management

We thank the NIH for this thoughtful Request for Information (RFI) on dbGaP Policies and Procedures. We commend dbGaP for furthering the cause of data sharing and appreciate its efforts to expedite the review process in the face of an exponentially growing volume of data submissions and data requests. As one of the largest U.S. genome centers, Broad Institute has had extensive experience with all aspects of dbGaP, ranging from data deposition to data discovery and retrieval. Moreover, members of our community have been involved in crafting the frameworks for data access and data use for initiatives such as such as the All of Us Research Program and the Global Alliance for Genomics and Health, and we have played a key role in data coordination for large-scale projects such as TCGA, 1000 Genomes, ExAC and GTEx. We are thus well-positioned to comment on the current state and future role of dbGaP. In crafting this response, we have attempted to draw upon these experiences and summarize the perspectives of the Institution at large. Our fundamental view is that the community desperately needs improved mechanisms to support data sharing. Given the social investment that has gone into generating the data currently stored in dbGaP, the paucity of investment into technologies for making it usable seems misguided. The ethical discussion surrounding data sharing needs to be reframed to recognize that there is an ethical obligation to make these data more easily accessible to responsible researchers. In this response, we offered a solution for each problem we have identified in numbers 1 through 3. We welcome the opportunity to work with NIH and the community to help implement these solutions. In particular, two general areas stand out to us as urgently in need of change: 1. Genomics researchers are beginning to make greater use of cloud services, yet dbGaP is not creating Trusted Partner or other mechanisms to allow for groups to host data on the cloud (with the exception of a few thoughtful programs such as the NCI Cloud Pilots). This creates the wasteful and unworkable situation where researchers must store multiple copies of the same dataset on the same cloud, as there is currently no mechanism to allow the researchers to access a common copy. 2. Enormous amounts of time are spent by Data Access Committees (DACs) policing whether a given data access request is consistent with a given data use restriction. This process is extraordinarily frustrating as there is a high degree of variability on granting access depending on the DAC. It is our view that datause restrictions can be structured as ontologies, so that this consistency checking can be automated by computers.
Were this done, researchers would be able to search for datasets that are consistent with their research purposes. The work of DACs would be substantially reduced, allowing for focus on of a minority of harder requests. We provided a solution to this problem.

Submission Date
04/07/2017
Name
Joel Hirschhorn
Primary Purpose of dbGaP Use
Data Access / Download
What is your level of experience with dbGaP?
Experienced (used many times over the course of several years)
Role/Other Role
Scientific Researcher
Type of Organization/Other Type
University
Name of Organization
Broad Institute, Boston Children's Hospital, Harvard Medical School Other Type of Organization

Information Requested

1. dbGaP Study Registration and Data Submission

2. dbGaP Data Access Request (DAR) and Review
I have a protocol that involves nearly 100 datasets from dbGAP, covered by multiple DACs. I would like to describe the difficulties and offer potential solutions that are relevant to my experience. Problem 1: The approval, amendment and renewal processes seem needlessly onerous and burdensome. The need for yearly renewal for the same science seems unjustified. Furthermore, amendments or renewals of the protocol are reviewed by all of the DACs, and the DACs are often slow and do not appear to be consistent or coordinated. Nearly every renewal or amendment results in one or more DACs requesting additional changes, sometimes in contradiction to requests from other DACs. IRB approval and sometimes even secondary layers of approval are often required for studying deidentified datasets, further adding to the complexity of gaining access to data. The net effect of this process is to dissuade many researchers from continuing, or even beginning, to invest time and effort in obtaining and maintaining access to dbGAP data. Although each action of each DAC is usually reasonable and defensible, the total effect is something akin to death by a thousand cuts. Potential solution 1a: Automate the approval and amendment process as much as possible (see comments from the Broad Institute). Potential solution 1b: Simplify the process of amendment and renewal with some or all of the following changes: (i) Eliminate the requirement for yearly renewal; instead request notification of publications that use dbGAP datasets. (ii) Allow addition of collaborators within the same institution without formal review by DACs, beyond simple and quick verification by a central body. (iii) Allow addition of cohorts without necessitating re-review by all DACs. Potential solution 1c: Have a single DAC make decisions for each protocol, or at least standardize and coordinate the approval and renewal process across DACs. Potential solution 1d: Establish guidelines, and work with contributing studies, to minimize the number of datasets that require explicit IRB approval or secondary review for access. Problem 2: dbGAP datasets are not consistently QCed, and phenotype data are often incomplete. Potential solution 2a: Fund one or more groups to establish a uniform pipeline for systematic QC of all genomic data in dbGAP, and to deposit the cleaned versions in dbGAP Potential solution 2b: Work with NIH program officers to require that if phenotype data are used in published association studies with genotype data that are covered by the data sharing policy, then the phenotype data should also be deposited in dbGAP. Require that the phenotype data be deposited in a standardized format (using terms from a phenotype ontology). Problem 3: There is unnecessary diversion of NIH funds when investigators each use grant funding to pay to store multiple copies of large dbGAP data sets. Potential solution 3: Allow trusted partners to store data in the cloud (see comments from Broad Institute).

3. Policies for the Management and Use of dbGaP Data

- **Alternate controlled-access models**
  I echo the comments of the Broad Institute.
- **Benefits and risks associated with the availability of genomic study summary statistics**
  I agree with the comments of Mark Daly. I believe harm has been done by promoting the potential risk of re-identification via summary statistics. Many powerful tools have been developed to derive important insights from summary statistics, especially when multiple phenotypes are considered simultaneously. The pace of discovery has been substantially slowed by limited access to these valuable summary statistics. Indeed, there is a strong ethical

commitment to research participants that is being downweighted by restricting access to summary statistics: the commitment to strive to make beneficial discoveries based on the selfless contributions of the participants. If there is no or minimal risk, then making these statistics inaccessible reneges on that commitment. Under nearly every circumstance, real world scenarios pose no or extremely low risk even if re-identification were possible, so the risk-benefit ratio of releasing these summary statistics should be favorable. Potential solution: NIH should rethink policies around summary statistics and the default should be public release of summary association statistics unless a clear case for harm can be made in specific circumstances. Even under those circumstances, summary statistics should be readily available to registered users after a simple and quick commitment not to use the statistics for reidentification (and perhaps also not to make inferences of population history for certain populations), without requiring burdensome application procedures.

- **NIH is interested in public comments on benefits and risks of such reference use of dbGaP data for research participants, patients, and the scientific community**

4. General Comments. NIH welcomes general comments on any other topics with regard to dbGaP data submission, access, and management

Submission Date
04/07/2017
Name
Angela Patterson
Primary Purpose of dbGaP Use
Study Registration / Data Submission
What is your level of experience with dbGaP?
Never used dbGaP
Role/Other Role
IRB operations
Type of Organization/Other Type
Health Care Delivery Organization
Name of Organization
Mayo Clinic

Information Requested

1. dbGaP Study Registration and Data Submission

2. dbGaP Data Access Request (DAR) and Review

3. Policies for the Management and Use of dbGaP Data

- **Alternate controlled-access models**
- **Benefits and risks associated with the availability of genomic study summary statistics**
- **NIH is interested in public comments on benefits and risks of such reference use of dbGaP data for research participants, patients, and the scientific community**

4. General Comments. NIH welcomes general comments on any other topics with regard to dbGaP data submission, access, and management
We understand the checkboxes on the Institutional Certification form page 1, distinguishing data “available through unrestricted or controlled-access”, refer to the type of database to which data will be submitted (i.e. an unrestricted or controlled access database), rather than the study participants’ consent (i.e. unrestricted consent or with data use limitations). If this is correct, how may IRBs confirm whether the relevant databases are controlled-access or unrestricted? The Data Repositories list on the GDS website includes a lot of information, but whether each repository is controlled-access or unrestricted is not readily identifiable. Do the various funding institutes at times have their own additional requirements? We received a request from one institute in the past to designate a particular “consent category” (from a list provided by the IC) in addition to the information provided on the standard Institutional Certification form. It would be helpful to either have a standard process across all ICs, or if there are differing requirements, a clear way to communicate those requirements to PIs and IRBs. Suggest the “Institutional Certifications” web page be linked from the menu on the left (Under “Home”, “Policy”, “Policy Oversight”, etc). It can take some digging to find the Institutional Certifications page. Please consider modifying the “Original Study Name” and “Project Title” text boxes on the Institutional Certification forms to allow longer titles to be readable on the completed form. The current fields cause text to get smaller for a longer title. Also, if the title length exceeds the length of the text box, text beyond the box length is not visible on the completed form. Because samples may be collected under multiple IRB projects, the “Original Study Name” field on the Institutional Certification forms should allow for multiple study titles. Also, please clarify the “Project Title for Data to be Submitted”, and how/whether this differs from the Original Study Name. GDS policy C.4, second paragraph, indicates that informed consent must be obtained from participants for future use of their data for studies initiated after 1/25/2015. However, GDS Section C.4, third paragraph, indicates that informed consent must be obtained for studies using data from specimens created or collected after 1/25/2015. Please reconcile these statements for consistency. Does the policy apply to studies initiated after 1/25/2015, or to samples created or collected after that date? E.g. does the policy apply to samples collected after 1/25/2015 under a study that initiated in 2014? Please expand the capacity of the “Data Use Limitation” fields on the Institutional Certification forms. Particularly if category “Other” is used, limitation descriptions can exceed the text box length, and text is then not visible on the completed form. The “NIH Guidance on Consent for Future Research Use and Broad Sharing of Human Genomic and Phenotypic Data Subject to the NIH Genomic Data Sharing Policy”, linked from the GDS FAQ, seems to be the clearest description of the information expected/required to be

in consent forms after 1/25/2015. Could this guidance be included within, or linked from, the GDS policy itself?

Submission Date
04/07/2017
Name
Aras Eftekhari
Primary Purpose of dbGaP Use
Both Study Registration / Data Submission and Data Access / Download
What is your level of experience with dbGaP?
Experienced (used many times over the course of several years)
Role/Other Role
Scientific Researchers, IT Consultants, Bioinformaticians
Type of Organization/Other Type
Management, IT, and Consulting Firm
Name of Organization
Attain, LLC

Information Requested

1. dbGaP Study Registration and Data Submission
dbGaP would benefit from leveraging a metadata-driven common data model using open standards, similar to other metadata repositories at HHS. This would allow for minimized harmonization and curation costs and would make reuse and reproduction of results more effective and accurate.

2. dbGaP Data Access Request (DAR) and Review
Regarding downloading data from dbGAP, some of the dataset are not stored at dbGAP any more like TCGA data, so the access is via dbGAP but the actual data and download is from other systems. The access authorization request information is passed to these detached portal every 24 hrs, but if not updated, the users' access is delayed. Users have to download and install software which helps in downloading (Aspera connect) and decrypt (SRAtoolkit) the actual scientific data. This process might be changed and improved using cloud technology – and GLOBUS would be a good choice to talk about here for data transfer.

3. Policies for the Management and Use of dbGaP Data

- **Alternate controlled-access models**
  The basic types of access control are: capability-based (access equals permission) (Access Control Level) ACL-based (each subunit [eg: file] keeps has a list of who can and cannot access it). Obviously these can be re-imagined and combined into different access control schemes (like the grey-list) but pretty much everything boils down to this. From a security perspective, a restrictive ACL-based permission system is probably the most secure. But if you aren't too concerned with security, you can always create categories for your ACL system and put rules in place around what is needed to access each subunit (file). From there, you can introduce roles as-needed instead of an individualized or all-or-nothing policy. Another way to improve responsiveness of access control would be to nominate community supervisors, who's job it is to manage who has access to a given role/group (so Ella Temprosa might manage who has access to COMETS data, even though she's not on the dbgap team). Correspondingly, this works best if there is a request and approval system in place for roles rather than an "email and update" type workflow. If you are outside of NIH, you need to be a PI or a local supervisor to apply for access, but an easier way was to be added as a downloader to a PI's access, which we did. So, there should be another way for groups with no PI or authorized supervisor to be able to apply for access. Sometimes the users mention there is no proper reason mentioned for application suspension. When users apply for access, it is by project for ex: TCGA, so when the user now needs access to another dataset for ex: TARGET, users state that they need to go through the process all over. An easier way will be to access previous records and add TARGET to it. The dbGAP user account is linked with eRA commons account, there is no other way for people with no eRA commons account, as far as I know to be able to get login access. Getting an eRA account assumes you are a US citizen and is harder for non-citizens, and your organization should be registered with eRA.
- **Benefits and risks associated with the availability of genomic study summary statistics**
  Risks of patient identification is always there with any kind of genomic data, but becomes more evident when this kind of data sits in different dispersed data repositories as there is more information available. Published theories like randomization and generalization can be applied to mitigate the risk of patient identification at the same time abiding by HIPAA regulations. Revisiting standards like Safe Harbor, which have known disadvantages when it comes to

collecting information required for genetic research versus deidentification, for example: most importantly genetic data is not considered important to be removed or generalized from records.

- **NIH is interested in public comments on benefits and risks of such reference use of dbGaP data for research participants, patients, and the scientific community**
  Pooling data from multiple repositories leads to a meaningful interpretation and diagnostics of the diease and the related data. Data from dbGAP certainly accelerates biomedical research, encourages collaboration and hence benefits and leads to better and more precise treatment options. Reference data from dbGAP can help find evidence to certain scientific facts and make informed decisions in treatment process. However, there might be data which could lead to conclusions that contrast with the existing knowledge. More clinical data and analysis is always helpful in such risky situations, as well. The Bionimbus Protected Data Cloud (PDC) is a secure biomedical cloud operated at FISMA moderate as IaaS with an NIH Trusted Partner status for analyzing and sharing protected datasets. The Bionimbus PDC is a collaboration between the University of Chicago Center for Data Intensive Science (CDIS) and the Open Commons Consortium (OCC). The Bionimbus PDC allows users authorized by NIH to compute over human genomic data in a secure compliant fashion.

4. General Comments. NIH welcomes general comments on any other topics with regard to dbGaP data submission, access, and management

General Comments- I would add to include more transparent documentation of the approval process. There has been a lot of confusion even on how to apply for access. Currently everything is on this page - https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?page=login A clear, transparent, easy to use (more intuitive) process.Is it possible to provide links to GDC which has some analysis capabilities. Would it be possible to use the GDC architecture within dbGap. (https://bionimbus-pdc.opensciencedatacloud.org/) Management- At Attain, we provide biomedical informatics support, management consulting, application development, and software engineering across HHS. We offer a unique, effective, and efficient project management model in the biomedical research space that utilizes Team Science and Interdisciplinary management concepts. It is currently in use and benefitng our federal clients at the following ICs and DOCs and we would be delighted to demonstrate our capabilities for dbGaP: - National Library of Medicine (NLM) - National Cancer Institute (NCI) - Center for Biomedical Informatics and Information Technology (CBIIT) - Clinical Evaluation Therapy Program (CTEP) - Center for Cancer Research (CCR) - National Institute for Environmental Health Sciences (NIEHS) - Big Data to Knowledge (BD2K) - Food and Drug Association (FDA) - Center for Disease Control (CDC) - National Center for Biotechnology Information (NCBI)

Submission Date
04/07/2017
Name
Daniel MacArthur
Primary Purpose of dbGaP Use
Both Study Registration / Data Submission and Data Access / Download
What is your level of experience with dbGaP?
Experienced (used many times over the course of several years)
Role/Other Role
Scientific Researcher
Type of Organization/Other Type
Nonprofit Research Organization
Name of Organization
Broad Institute of MIT and Harvard

Information Requested

1. dbGaP Study Registration and Data Submission
The study registration and submission process is overly complex, manual, lacks a systematic manner in which data is curated, and does not accurately capture critical phenotypic information. Proposed solutions: 1. Automate the sample registration process. Much of the sample registration process is manual and overly time consuming. Currently, this registration requires communicating directly with a dbGaP officer, which is not scalable. 2. Develop simple tools for the deposition of phenotypic data. Phenotypes are not structured on deposition, greatly complicating meaningful secondary use. NIH must either develop or encourage others to develop tools for structuring phenotypes. These structured phenotypes must be part of a comprehensive ontology system to enable flexible queries and searching. 3. Require deposition of all phenotypic data that is published. Again, meaningful secondary use requires good phenotypic data. If investigators are funded by NIH and have the phenotypic data and the tools are provided for easy deposition (#2 above), NIH should require the deposition of the data (at minimal, the phenotypic data that was already included in a publication). 4. Data Submission and Curation.
There exist several groups with expertise in large-scale data engineering, as a result of efforts such as TCGA, GTEx, ExAC and TOPMed. This expertise should be leveraged by creating a small number of centers (but more than one), that are charged with a. Processing of genomic data with various best practices pipelines. This mandate should include reprocessing data with new builds of the genome, or when there is an algorithmic advance in variant calling significant enough to justify it. b. Mapping phenotypes to ontologies. c. Structuring data use restrictions so that they are mapped to a standard ontology (see Section 2, below). 5. Computing Infrastructure: The curated datasets should be placed onto public clouds for use by the community (more than one is crucial to maintain competition), in addition to NIH-owned servers from which it can be downloaded. It is crucial that these datasets be directly accessible by users, and that users not be required to go through any given platform to access a given dataset. 6. Access Control: dbGaP should continue to adjudicate access to publicly funded datasets. Under this model, dbGaP would continue to be an identity provider that adjudicates who is a “trusted researcher,” and researchers would continue to apply to this body to receive permissions to access various datasets. The data should be stored on one or more cloud infrastructures; as access is granted, this would be mirrored across infrastructures. dbGaP should make whitelists of which researchers are allowed to access the various applications publicly available via APIs so that third-party applications could respect these permissions. In summary, the above model would facilitate access: a. By ceding data curation to groups with expertise in it, this work would no longer fall squarely on the shoulders of dbGaP. b. By making data use restrictions machine-readable, the work of DACs would be greatly simplified. c. By utilizing clouds, there would be less need for downloading data to new environments, which significantly taxes the current system.

2. dbGaP Data Access Request (DAR) and Review
The current system for granting data access is entirely manual. This process is both: (a) highly variable based on the Data Access Committee; (b) simply not scalable. Additionally, the annual renewal and reporting process is extraordinarily burdensome for investigators to complete and NIH staff to review and provides little benefit. Solutions: I. Standardize Data Use Letters (DULs) and Data Access Requests (DARs). The Broad has developed a user friendly tool and has piloted this experiment. My colleagues at the Broad Institute have inspected a collection of nearly 132 DULs at the institute. From that exercise, they determined that 95% could be structured into an ontology that contained the following 5 main categories (see Figure below): i) disease- specific restrictions, ii) commercial restrictions, iii) restrictions to special populations, iv) restrictions on research use, v) General

Research Use. If DULs were structured to follow a standard ontology, researchers would be able to search for datasets consistent with their research purpose. It would also greatly facilitate the work of the DAC, as the effort of checking whether a DAR is consistent with a DUL would be automatable and thus scalable. Towards this end, the Broad Institute has developed an open-source software package ("DUOS") for structuring DULs and filtering them by research purpose. This software package also contains interfaces for DACs to facilitate review of data access requests. We would happily donate this software to dbGaP for use by its DACs. II. Phenotype-based search Just as the previous example noted the need to data-use-enabled search, there is a significant need for phenotype- based search. Researchers would greatly benefit from the ability to search for all samples with a given phenotype in constructing cohorts. This search should be ontology-aware (e.g., search for "cancer" includes "angiosarcoma" etc'). III. Extending and Renewing Data Access Requests Currently, if a researcher has approval to study one dataset and later wants to utilize a newly deposited dataset for the same research, this entails an entirely new application. There should be a mechanism by which researchers can extend an already-approved application to include additional datasets as they emerge. Finally, the annual renewal process is unnecessarily burdensome for both Investigators and NIH staff. We propose that all investigators agree to a code of conduct rather than annual renewal and that an audit function by the "trusted partner" replace annual renewals. IV. Better coordination of DARs with Cohorts Many of the traditional large-scale cohorts are governed by an added layer of access control--either a local IRB or a DAC that is affiliated with the cohort. This can greatly complicate and prolong the process of accessing data, and we have experienced inconsistency between cohorts with regard to protocols and policies. We propose that this additional layer be streamlined where one representative from each cohort is designated to make time-limited decisions. If there are additional limitation on data use, the group should make them transparent so that researchers will only ask for the data if they can meet any of the additional layers of review.

3. Policies for the Management and Use of dbGaP Data

- **Alternate controlled-access models**
  dbGaP presumes that researchers will download data to their own infrastructure. Accessibility is a significant limitation, as many investigators do not have access to the compute and storage infrastructure to host datasets of large scale. Instead, many groups have started to use cloud services. There is, however, no cloud-based system that allows for a single point of storage. This creates the wasteful and unworkable situation where researchers must store multiple copies of the same dataset on the same cloud, as there is currently no mechanism to allow the researchers to access a common copy. A key solution to this problem is to 'Bring researchers to the data' through the creation of a pathway for institutions to achieve Trusted Partner status, especially with regards to the utilization of cloud services A number of platforms have emerged to address this challenge and host data on public clouds. However, the community is blocked from using them as there is, in general, no pathway to becoming a trusted partner to distribute data; this is an unnecessary regulatory obstacle that can and should be readily remedied. As a concrete example of how wasteful the current approach is, the NCI has invested significant resources into funding the creation of three "Cloud Pilots," which host TCGA data via a cloud-based platforms (Broad has received one of these awards). Each has Authority To Operate (ATO) as a FISMA Moderate environment, and each is integrated with dbGaP procedures such that, when the DAC grants access to a researcher to utilize TCGA data, it is mirrored in the access control of the cloud pilots. The Broad Institute and others have repeatedly asked for Trusted Partner status to host not only TCGA data through this platform, but also additional datasets that researchers in our community have already placed onto the cloud. Doing this would require no additional funding from the NIH, and it would resolve the current predicament of researchers who must store multiple copies of the same dataset on the same cloud, as there is currently no mechanism to allow the researchers to access a common copy. We have been unable to advance this conversation. The lack of an ability to formally apply for an receive Trusted Partner status thus results in a complete, curated copy of all Broad-generated data being present on the cloud, fully funded, to which no external individual has the permission to gain access. The lack of any progress on this issue represents a dramatic waste of NIH resources and scientific capabilities.
- **Benefits and risks associated with the availability of genomic study summary statistics**
  As the leader of the ExAC (Exome Aggregation Consortium) and gnomAD (Genome Aggregation Database) consortia, I am a proponent of the responsible release of genomic summary statistics, but here I speak for myself and not those consortia. ExAC and gnomAD are the leading providers of aggregate allele frequency data today, and that aggregate data is completely open to the public via our browsers; these projects were approved and classified as Not Human Subjects Research by our local IRB. In my experience, there have been huge benefits from the release of aggregate frequency data to the public, and extremely low risk to participants. ExAC and gnomAD are widely used, not only for basic science, but by clinical geneticists helping patients and by pharmaceutical scientists developing drugs, with over 7.7 million page views since 2014 (3.3M over the last 12 months, by 137,000 unique users from 176 countries). If ExAC and gnomAD's summary statistics were only made available through a dbGaP-like controlled access system, it would only have a tiny fraction of those users, and society would only have a tiny fraction of the benefits. The main risks

associated with these variant frequency databases is that of reidentification and loss of privacy - but these risks are extremely small, as they require a technically sophisticated user who already possesses the genetic data of the participant in question. More importantly, because ExAC and gnomAD samples span so many phenotypes (cases and controls for many different adult-onset diseases) our theoretical hacker learns almost no useful information by inferring that an individual is part of the ExAC data set. No known privacy breach or other harm has ever come to any ExAC or gnomAD participant due to their data's aggregation and release. When it comes to the public release of allele frequency summary statistics, the risks are exceedingly small and theoretical, while the benefits are enormous and real. This remains true, in almost all cases, for other types of genomic summary statistic (such as case/control P value) as well; the only obvious potential exceptions to this are in the case of especially vulnerable populations such as case-only studies of drug users or HIV- positive individuals, for whom the inference of belonging to a study carries the risk of real-world harm. Therefore, I urge the NIH to formally reclassify genomic summary statistics and other aggregate measures such as allele frequency as open access, with the exception of case-only data sets from particularly vulnerable populations, and explicitly make them freely available outside the dbGaP controlled access system.

- **NIH is interested in public comments on benefits and risks of such reference use of dbGaP data for research participants, patients, and the scientific community**

4. General Comments. NIH welcomes general comments on any other topics with regard to dbGaP data submission, access, and management

As the coordinator of the ExAC (Exome Aggregation Consortium) and gnomAD (Genome Aggregation Database) consortia, I have had considerable experience with the dbGaP Data Access Request and Review process since 2014; here I speak for myself and not those consortia. In the past three years, I have applied for over 30 exome and genome datasets for inclusion and release in aggregate form in the ExAC and gnomAD databases. While I appreciate the hard work and good intentions of dbGaP staff, this process has revealed major shortcomings of the system. Several proximal problems desperately need to be addressed: - Data providers are permitted to add extra requirements to applications that restrict data sharing without adding any benefit or reducing any risk to participants. Examples include requirements for local IRB approval, custom additional forms, and letters of collaboration.

- The yearly renewal process adds administrative burden without benefit to participants. We suggest removing this and replacing with a simple code of user conduct.
- Perversely, there is no mechanism to add additional datasets to an existing project without reapplying for every previously approved dataset within a project - an approach that adds extra burdens to DACs for no good reason, and which is a particular problem for data aggregation efforts like ExAC.
- There is little consistency between data access committees, or from committees over time. For instance, we recently had several previously approved requests suddenly denied because the composition of the DAC in question had changed.
- The website is overburdened, and frequently unavailable without notice. Moving forward, the fundamental problem faced by dbGaP is scalability. The datasets that are desired (and often required) to be deposited in dbGaP already exceed its storage capacity and strain its design. These problems will grow as NIH-funded WGS data production increases. To futureproof controlled-access data sharing, dbGaP needs a cloud-based solution for controlled access data that allows for data storage for all NIH- funded sequence projects, and reduces (often NIH-funded) costs by eliminating duplicate data storage. In this system, scientists would come to the data instead of bringing the data to each scientist. An optimal controlled-access data sharing system would bring scalability to every step in the review process. Each project would be reviewed once, by a single Data Access Committee, and all datasets with Data Use Restrictions compatible with that project would be approved at once - future datasets would be automatically approved for the project as they are uploaded to the system. The Data Use Requests and Data Use Letters would be standardized so as to allow for automated matching. Scientists could then analyze all approved datasets for their project, alone or in combination with private data, in the cloud without ever needing to download the datasets locally. Large-scale data providers should be able to apply for Trusted Partner status to redistribute data to approved researchers. Such a system would serve the needs of the genomics community for the next decade of rapid advances in genomic data production and reuse.