Data Protection and Access in NHLBI DataSTAGE: FAQs

1. **What is DataSTAGE?**

   The NHLBI-supported DataSTAGE (Storage, Toolspace, Access and analytics for biG data Empowerment) (www.nhlbidatastage.org) is a cloud-based infrastructure where heart, lung, blood, and sleep researchers can go to find, search, access, share, cross-link, and compute on large scale datasets. It provides tools, applications, and workflows to enable those capabilities in secure workspaces.

2. **How is data secured in DataSTAGE?**

   DataSTAGE operates on Amazon Web Services (AWS) and Google Cloud Platform (GCP). Both AWS and GCP have received an Authority to Operate (ATO) from the General Services Administration FedRAMP (www.fedramp.gov) following a rigorous assessment process by a third party assessor. The NHLBI Chief Information Officer has reviewed the System Security Plan from each system that comprises the DataSTAGE environment and has issued an ATO for their system to operate at the Moderate level. This type of authorization is consistent with the National Institute of Standards and Technology guidance and complies with all requirements of the Federal Information Security Management Act.

3. **Who manages the data in the DataSTAGE cloud environments?**

   Similar to dbGaP, the data within DataSTAGE is managed by NHLBI staff and contractors. All staff and contractors with access to the data hold a Public Trust Clearance that is based on an extensive background check. All activities including data access are logged and monitored.

4. **How is access granted to controlled-access data in DataSTAGE?**

   Access to data is controlled by the NHLBI Data Access Committee (DAC) utilizing the database of Genotypes and Phenotypes (dbGaP) permissions infrastructure and the NHLBI Data Access Committee (DAC). In order to access controlled-access data in DataSTAGE, a user must have an approved Data Access Request (DAR) in dbGaP (see also FAQs 8 and 9). The role of the NHLBI DAC is to review and approve (or deny) investigator submitted DARs and to ensure investigator compliance with the NIH Genomic Data Sharing Policy.

5. **How does an investigator request access to controlled-access data in DataSTAGE?**

   A user must follow the normative dbGaP data access request process by submitting a DAR. Once the DAR has been approved in dbGaP by the NHLBI DAC, the data will be available to them in DataSTAGE.

6. **Can someone gain access to controlled-access data without an approved dbGaP Data Access Request (DAR)?**

   No, the only way to gain access to controlled access data is via the dbGaP DAR process, (see also FAQs 8 and 9).

7. **How does DataSTAGE reduce the potential risk of data being used for unintended purposes compared to dbGaP?**

   Access to data that is placed in dbGaP is controlled via the Data Access Request (DAR) process. Traditionally, after approval, a user downloads the data from dbGaP and there is no further direct control or monitoring of how the data is used or potentially redistributed. Within DataSTAGE, while access is managed in the same way as dbGaP, the data cannot be downloaded from DataSTAGE and all use of the data is logged and monitored. Therefore, the risk of unintended use of the data is reduced in DataSTAGE.

8. **Is the NHLBI pursuing novel methods of access to controlled-access data?**

   The NHLBI is pursuing data access processes and tools that could provide access to data in a more facile manner in compliance with NIH Policy and the Genomic Data User Code of Conduct, participant informed consents, and with equivalent or better security than dbGaP. The NHLBI aims to improve and streamline the data access process by applying technology to ease the burden on investigators and speed access while providing a secure environment with protections against misuse of data.

9. **What are some of the novel methods of access that NHLBI is pursuing?**

   Library Card/Data Passport

   The Library Card[1] model is designed to streamline the process for requesting access to data by allowing an Institutional Signing Official to provide a general approval or authorization for an investigator to apply for data for a period of time, rather than having to approve multiple DARs. Investigators will still have to complete a DAR for each project, but the Library Card will give them pre-approval from their Institutional Signing Official. Also, the DAR will still need to be approved by the DAC before access to data is given. The NHLBI is working with NHGRI and the Broad Institute to conduct pilots of the Library Card concept.

   Data Use Ontology System (DUOS)

   The Data Use Ontology System (DUOS) is an implementation of the Global Alliance for Genomics and Health (GA4GH) Data Use Ontology that supports machine readable encoding of Data Use Limitations (DUL) for a data set and the Research Use Statement within a DAR. DUOS can facilitate and streamline the evaluation of data access requests and could potentially provide automated approval of access requests in cases where

[1] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5851345/

non-sensitive data is being requested. The NHLBI is experimenting with DUOS as a tool to support the NHLBI DAC.

10. Does DataSTAGE follow the TOPMed timetable for release of TOPMed data?

Yes, the same policies for release of data used by TOPMed are enforced by DataSTAGE.

11. Does DataSTAGE enforce the requirement for IRB approval and other Data Use Limitations?

Because DataSTAGE uses the dbGaP permission infrastructure for controlling access to data, the same requirements for IRB approval or other Data Use Limitations are enforced. Should DataSTAGE implement novel methods for requesting and approving access, these same requirements will be enforced.

12. If my study has special requirements for access to data such as a Letter of Collaboration, will DataSTAGE enforce this requirement?

Because DataSTAGE uses dbGaP permission infrastructure for controlling access to data, any additional requirements for approval are enforced. Should DataSTAGE implement novel methods for requesting and approving access, these same requirements will be enforced. Study PIs for studies with special requirements or especially sensitive data are encouraged to contact the NHLBI DataSTAGE (nhlbidatastage@mail.nih.gov) and NHLBI DAC (nhlbigeneticdata@mail.nih.gov) project office teams to discuss any concerns.

13. As a data submitter, do I need to do anything to move my data from dbGaP to DataSTAGE? Do I need to submit another institutional certification?

No, dbGaP and DataSTAGE staff will ensure data is safely available through DataSTAGE. Pursuant to NIH policy on sharing of research data (NOT-OD-03-032) and the NIH Genomic Data Sharing Policy (NOT-OD-14-124 and NOT-OD-19-023), the NHLBI has classified DataSTAGE as an NIH-designated repository and so a new certification is not needed.

14. Can my institution or I opt out of having my study's data in DataSTAGE?

Pursuant to NIH policy on sharing of research data (NOT-OD-03-032) and the NIH Genomic Data Sharing Policy (NOT-OD-14-124 and NOT-OD-19-023), the NHLBI has classified DataSTAGE as an NIH-designated repository. Therefore, in general a PI or Institution cannot opt out of having their study's data in DataSTAGE. Under certain circumstances where there is a specific sensitivity or risk, investigators are encouraged to discuss any concerns with the NHLBI DataSTAGE team and/or NHLBI DAC.

15. Do I need to change anything in my Data Access Request in order to use DataSTAGE?

You may need to add a new Cloud Use Statement as a component of your DAR that specifically references the NHLBI DataSTAGE as the environment to be used. A sample Cloud Use Statement can be found here (below).

16. I've heard that with DataSTAGE there is the ability to query data and see what data is available before a Data Access Request is submitted. How does this align with controlled access?

The NIH Genomic Data Sharing policy, via its guidance on Genomic Summary Results, encourages minimal barriers to access genomic summary results. Therefore, investigators will be able to see and query data at the summary level within DataSTAGE without an approved DAR. Logging in via an authorized account is required before a user can query the data. This will enable logging and monitoring of users and provide an opportunity to remind users of the three principles for responsible research use (no attempt to re-identify, use only for research or health purposes, and review of the responsible genomic data use informational materials). In addition, minimum cell size restrictions, which are common across databases, will help protect the confidentiality of individuals.

NHLBI DataSTAGE Cloud Use Statement:

The NHLBI-supported DataSTAGE (Storage, Toolspace, Access and analytics for biG data Empowerment) (www.nhlbidatastage.org) is a cloud-based infrastructure where heart, lung, blood, and sleep (HLBS) researchers can go to find, search, access, share, cross-link, and compute on large scale datasets. It will provide tools, applications, and workflows to enable those capabilities in secure workspaces.

The DataSTAGE will employ Amazon Web Services and Google Cloud Platform for data storage and compute. DataSTAGE comprises the Data Commons Framework Services (DCFS) hosted and operated by the University of Chicago. DCFS will provide the gold master data reference as well as authorization/authentication and indexing services. The DCFS will also enable security interoperability with the secure workspaces. Workspaces will be provided by Terra, hosted and operated by the Broad Institute; Fair4Cures, hosted and operated by Seven Bridges Genomics; and i2b2/tranSMART, hosted by University of Chicago and operated by Harvard Medical School.

For the NHLBI DataSTAGE, the NHLBI Designated Authorizing Official has recognized the Authority to Operate (ATO) issued to the Broad Institute, University of Chicago and Seven Bridges Genomics as presenting acceptable risk, and therefore the NCI ATO serves as an Interim Authority to Test (IATT) when used by designated TOPMed investigators and collaborators.

Amazon Web Services (AWS) is a secure cloud services platform offering compute power, database storage, content delivery and other functionality that will allow us to deploy sophisticated analysis efforts on large scale phenotypic and genomic datasets quickly and cost-effectively. It is a secure, durable technology platform with industry-recognized certifications and audits: PCI DSS Level 1, ISO 27001, FISMA Moderate, FedRAMP, HIPAA, and SOC 1 (formerly referred to as SAS 70 and/or SSAE 16) and SOC 2 audit reports. Their services and data centers have multiple layers of operational and physical security to ensure the integrity and safety of data. AWS has summarized how their platform supports compliance with controlled-access datasets in a white paper, including best practices for dbGaP: https://d0.awsstatic.com/whitepapers/compliance/AWS_dBGaP_Genomics_on_AWS_Best_Practices.pdf

Google Cloud Platform is a cloud computing service by Google that offers hosting on the same supporting infrastructure that Google uses internally for end-user products like Google Search. Google undergoes several independent third party audits on a regular basis to provide verification of security, privacy and compliance controls including annual audits for SSAE 16/ISAE 3402 Type II. Google's infrastructure provides reliable information security that can meet or exceed the requirements of HIPAA and protected health information. The Google Cloud Platform has summarized its services with respect to genomics data processing in a white paper here: https://cloud.google.com/genomics/resources/google-genomics-whitepaper.pdf