|  |  |
| --- | --- |
| **For PHS Use Only** | |
| **Privacy Board Approval Date:** |  |
| **Part D Approval Date:** |  |

|  |  |
| --- | --- |
| **DUA Signatory name and title** | Brian Del Bono, Sr. Contract & Grant Officer, Office of Sponsored Research |
| **Requesting Organization** | Board of Trustees of the Leland Stanford Junior University (Stanford University) |
| **Type of Organization** | Academic |
| **Study PI** (if different from DUA User) | Mark Cullen, MD |
| **Study Title** | Protocol 36332 – Secure Academic Data Commons: Data Receipt, De-identification, Risk Assessment, Storage, Curation and Access Control |
| **Date of IRB Approval** | May 28, 2019 |

## DATA MANAGEMENT PLAN

# PHYSICAL POSSESSION AND STORAGE OF PHS DATA FILES

* 1. Responsibility for technical infrastructure, data security and data curation including name(s) and job title(s).

Mark Cullen, Director, Stanford Center for Population Health Sciences

Ruth Marinshaw, Director of Research Computing, Stanford University

Somalee Datta, Director of Research IT, IRT Research Technology, School of Medicine, Stanford

Susan Weber, Director of Engineering, IRT, IRT Research Technology, School of Medicine, Stanford

Isabella Chu, Associate Director, Data Core Stanford Center for Population Health Sciences

Lesley Park, Associate Director, Data Core Stanford Center for Population Health Sciences

Valerie Meausoone, Assistant Director, Data Core Stanford Center for Population Health Sciences

Erika Tribett, Project Manager, Stanford Center for Population Health Sciences

Emma Hallgren, Data Manager, Stanford Center for Population Health Sciences

The list above represents the management of Stanford’s research computing staff and the Stanford Center for Population Health Sciences Data Core. All individuals with custodial responsibility for the data environment and administration will be either listed above will report to one of these individuals.

* 1. Maintenance of data inventory.

All research at Stanford using data hosted by PHS will be conducted on an Stanford Information Security Office (ISO) reviewed and approved computational environment data that abides by Stanford’s computer and network usage policy (http://adminguide.stanford.edu/62.pdf) and information security policy (http://adminguide.stanford.edu/63.pdf), as well as minimum security standards (https://itservices.stanford.edu/guide/securitystandards) for computational environments for moderate and high risk data. All researchers working on data hosted by PHS must abide by the security and regulatory restrictions of both the data owner and PHS, whichever is more restrictive. For each computational environment, there must be a complete Data Risk Assessment which stipulates the risk profile of proposed datasets and verifies security standards which meet or exceed Stanford standards for storage and computation for the risk profile assigned to the dataset.

PHS tools for data ingestion log: a) the database name, b) the date that the data were received, c) the identity of the analyst who was is responsible for the curation of the data and d) the filenames along with e) their file types and sizes; f) any transformation or cleaning processes. Meta-data are then loaded onto our data portal and permissions are set so that potential users of data can view the presence of the data, data description, variable list and data documentation, deidentified data and raw data. Each of these tiers of data access (view existence of dataset, view meta-data and data documentation, view and use data) have requirements and permissions set at a granular level according to the data risk classification and Data Use Agreement with the data owner.

In the event data are destroyed, we will record: a) Data Destruction date and b) we will include a Certificate of Disposition including the date we submitted the Certification of Disposition to the owner of the data.

* 1. Verification of privacy and security trainings prior to use of research data.

PHS data are only shared with individuals who: a) have completed required Data Security, HIPAA and human subject protection (CITI) training(s) as applicable b) have had all electronic devices which may potentially be used to access the data or analytic datasets encrypted c) provided proof of this this encryption either via the institutional encryption tracking and verification system or a signed statement from the Data Security Office of the researcher’s institution d) have institutional review board (IRB) approval for their study or are included on an IRB approved study or determination, e) have a “need to know” status with regard to the data and cannot practically work on the project without it and; f) signed a Data Use Agreement (DUA) with Stanford PHS stating that they will only use the data for the stated and IRB approved research purposes and that they cannot share the data with any third party, download the data and will abide by all regulatory and security requirements for that dataset. In some cases it will be necessary to complete special trainings and separate approvals as required by the data owner. These will be tracked in the Stanford PHS data portal within the users account.

In addition to the conditions above, the signed Data Use Agreement will stipulate that only output results from analyses can be downloaded – that is, the researcher agrees never to download data or subsets of those/analysis files or publish a formula which could be used to derive a cell with fewer than 11 individuals and that all data outputs will conform to the data owners or Stanford PHS cell size policies (no cell smaller than 11), whichever is more restrictive. In rare instances, data download is permitted, however, individuals downloading data or analytic datasets must receive prior written approval from PHS and additional encryption and security verification.

File transfer activity is audited on an ongoing basis. In the event of a suspected or reported file download, the researcher in question will be contacted. In the event any PHS data has been downloaded to a computer or other device, the investigator will be required to purge the files from their computer or device. Disciplinary action up to and including termination of data access, termination of enrollment or employment, or other sanctions will bean taken as appropriate. All computers used to access PHS data must be encrypted and password protected as described above.

For individuals who have terminated affiliation with Stanford or Stanford research projects, termination of all access and permissions is handled centrally by HR and occurs on last day of work or enrollment. As a safeguard, for students and visiting scholars, requirements are set to expire in a timeframe which coincides with their tenure at Stanford. In some circumstances, a research collaborator may retain access to data for project completion or research collaborations. Such individuals must meet all the same training, encryption and security requirements as outlined above. For individuals terminating affiliation with Stanford and Stanford research projects, access to the data will be terminated the last day of that individual’s participation in the project unless continued access has been arranged in advance and all required security and regulatory steps have been completed. If required, they must also complete an addendum to Stanford’s agreement with the data owner to allow completion of the proposed research project.

* 1. Notification of project staffing changes.

The Principal Investigator (PI) of any research project using PHS data will promptly inform the Data Core via email of any staffing changes. The PI or their designee must also update the appropriate study team members within the Stanford PHS Data Portal. The Data Core will request an update of team members at least annually. The Data Core Manager will confirm access to data has been terminated for any individuals who should no longer have access. The PHS data portal is the preferred method of transmitting this type of information as it provides the notice in writing, is a searchable, indexed format, includes a date and time stamp and allows the sender to confirm receipt. Additionally, the PHS Data Portal tracks permissions on an ongoing basis and individuals who have a requirement drop out of compliance will have data access suspended.

* 1. Training and education programs for high risk data.

As stated in Section 1.3, access to data is only shared with individuals who have completed the necessary regulatory, security and training requirements. PHS has two trainings for data security: The first for high risk data containing PHI or PII, or derived from data containing PHI and PII and the second for data which has been designated moderate risk (eg, deidentified survey data).

* 1. Infrastructure (facilities, hardware, software, other) to secure high risk PHS data.

The PHS data infrastructure is explained in more detail in section 2.5. Data will be received by the Stanford PHS Data Core via secure FTP transfer or encrypted drives (whichever method is preferred by the data owner). All research using PHS data will be conducted on an ISO reviewed and approved secure computational environment that abides by Stanford’s computer and network usage policy (http://adminguide.stanford.edu/62.pdf) and information security policy (http://adminguide.stanford.edu/63.pdf), as well as minimum security standards (https://itservices.stanford.edu/guide/securitystandards) for computational environments for PHI and other high risk data.

The requirements associated with standards for high risk data include patching, server hosting in a secure data center, centralized logging, two-factor authentication, vulnerability scanning and mitigation, and intrusion detection. Intrusion detection practices are in place, in accordance with Stanford’s minimum security requirements. Data are replicated nightly to a replica system. Both the regular and replica servers adhere to all of the same security requirements. Stanford on premise servers are housed in a secure research data center, the Stanford Research Computing Facility, with keycard access limited only to authorized individuals. In addition, the server rack itself is locked; the key is only available to the system administrators of the server and the Data Center Manager. The Data Center has 28 cameras that capture motion 24x7. In case of facility power failure, supplemental power is provided by a generator. Testing of the ability to switch between commercial and generator power occurs monthly. The server has also been outfitted with self-encrypting hard drives. All statistical software necessary for analyses live on the server so that it is never necessary to download data onto local computers or other electronic devices. In addition to the above safeguards, server system administrators are required to attend two days of information security training each year and they only access the server from bastion host systems specifically secured by the Stanford Information Security Office.

Authorized system administrators can physically access the system components when needed for maintenance activities.

The Stanford University Network Access Control (SUNAC) service (https://uit.stanford.edu/service/sunac) and the Stanford PHS data management platform allow granular and configurable control of data access. The System Administrator (or Custodian of the data) are able to control remote access to data and other resources located behind University IT-managed firewalls. Using Workgroup Manager and Redivis (a platform which allows granular permissioning for data access), access can be granted and customized for any PHS Member and dataset by SUNet ID. The PHS Data Core is responsible for this permissioning process for the PHS data projects.

Stanford PHS requires encryption of all devices used to access Stanford high risk data resources—whether the computers or devices are owned personally or by the University. These standards apply to all computers used to access PHS data regardless of the school or institutional affiliation of the PHS member. The encryption status of every Stanford computer and device used to access Stanford systems (including the server where PHS data are stored) is tracked in an internal system and audited continuously. Each person accessing Stanford systems must fill out a data attestation form and attest to the agreement annually or whenever there is a change in either their devices or the types of data they access, whichever is most frequent. Each computer or device used to access PHI or other high risk data must have whole disk encryption. This policy is available here: https://med.stanford.edu/irt/security/information/policies.html.

Computer encryption for Stanford is conducted and verified using the Stanford Whole Disk Encryption (SWDE) service. The SWDE service is for both Windows and Macintosh computers that support native encryption. Once installed, all files are automatically encrypted. The data are protected while the computer is in standby or hibernation mode. This requirement applies to Stanford and personally owned computers or computers that are used by Stanford research collaborators that are used for research activities on the Stanford network or any Stanford system. In addition to SWDE, SUNet authentication is required to access Stanford data or systems. Personal computers require a 10 character PIN to open the computer. Additionally, all systems used to access data require two step authentication with a password. Passwords must comply with Stanford password guidelines (https://uit.stanford.edu/service/accounts/passwords). Passwords must be a minimum of 8 characters and passwords less than 12 characters require a combination of mixed case letters, numbers and special characters.

In the event of loss or theft of a computer with High Risk Data, Stanford policy requires notification of the Information Security Office (ISO). ISO in turn will use the logs to determine if a lost or stolen computer is a "reportable" event, possibly requiring notification of data owners or individuals within a dataset. Although we adhere to reporting requirements, PHS does not permit individuals to download any PHS data onto individual computers or devices. All analyses must be conducted in the secure PHS computational environment and researchers are only permitted to download outputs which comply to PHS cell size restrictions (no cell smaller than 11), or the cell size restrictions of the owner of the data whichever is more restrictive. Printing in the server environment is disabled. Thus, individuals are unable to print PHS data. In rare instances, if special circumstances require downloading a subset of data, encryption of the proposed device will be verified and documentation of this permission will be included in the researchers account on the PHS data portal.

* 1. Policies and procedures regarding the physical possession and storage of PHS data files.

The PHS data infrastructure is explained in more detail in section 2.5. Data will be received by the Stanford PHS Data Core via secure FTP transfer or encrypted drives (whichever method is preferred by the data owner). If the data owner chooses to transmit data using physical drives or discs, the drives, discs or media containing the original PHS files, as received, will be stored in a safe in a locked room accessible only to the PHS Data Core management. The combination of the safe, a SentrySafe SFW205CWB Water-Resistant Combination Safe, 2X-Large, is only known to the Data Core management team. Further, the data will not be physically moved or transmitted in any way from Stanford without written approval from PHS and a DUA and DMP that explicitly prescribes the uses and disposition of the data.

During transfer, drives will be accompanied by at least two members of either the PHS Data Core or the team responsible for the computational environment on which the data will reside. This chain of command will be documented and included in the records for that dataset.

* 1. System to track the status and roles of the research team.

The Stanford PHS Data Core, keeps records of study personnel and status with regard to protocols, data access and current Data Security, HIPAA and human subject protection (CITI) training(s) as applicable. This information can be verified in the Stanford e-protocol system, STARS and internal records kept by PHS. In this system roles are delineated and records are updated as protocol or personnel changes occur or on an annual basis, whichever is more frequent. In addition to listing relevant personnel on IRB protocols, the e-protocol system also confirms that required HIPAA and human subject protection (CITI) training(s) are completed and up to date. In the event research projects are completed, the Stanford Data Manger records the completion of the project and the access to data is terminated. The closure is also noted in the e-protocol system (where applicable) or on the PHS Data Portal for studies which are covered under umbrella protocols.

* 1. Physical and technical safeguards used to protect PHS data files (including physical access and logical access to the files).

These safeguards are also outlined in sections 1.6 and 2.5. The Stanford on prem computational environment abides by Stanford’s computer and network usage policy (http://adminguide.stanford.edu/62.pdf) and information security policy (http://adminguide.stanford.edu/63.pdf), as well as minimum security standards for servers (https://itservices.stanford.edu/guide/securitystandards) for PHI and other sensitive data.

The system is accessed via SSH, X2Go via SSH and RDP via the campus VPN. All authorized users of the data will connect to the Stanford network via VPN and will authenticate with two step authentication (SUNet and password with Duo or Text Code secondary authentication); they will then also authenticate to the server with SUNet and password.

Authorized system administrators can physically access the system components when needed for maintenance activities.

Individuals are not permitted to download any PHS data onto individual computers except in rare instances and with written permission. All analyses must be conducted in the approved server environment. Researchers are only permitted to download outputs which comply to PHS cell size restrictions, or the cell size restrictions of the owner of the data, whichever is more restrictive. PHS will audit data download activity on an ongoing basis. Individuals suspected of having PHS data will be notified and, in the event data have been downloaded, the individual will be required to wipe the data from their computer and further access to PHS data will be terminated. Disciplinary and corrective action will be taken as appropriate. Printing in the server environment is disabled. Thus, individuals are unable to print PHS data. Systems used to store PHS data must meet or exceed Stanford standards for storage and computation for the risk profile assigned to the project data.

1. **DATA SHARING, ELECTRONIC TRANSMISSION, DISTRIBUTION**
   1. Compliance with [NIH Data Sharing Policies](https://grants.nih.gov/grants/NIH-Public-Access-Plan.pdf).

The NIH, and increasingly other funders, require that data from all research funded with federal or foundation dollars be made discoverable and available for verification of findings, replication and follow on research. The PHS data portal allows investigators to post their data and code in a discoverable, searchable format. Individuals requesting dataset access will have to complete the appropriate requirements (as outlined in Section 1.3) assigned to the dataset according to risk classification and any proprietary restrictions the dataset may carry. Data access and activity will be logged and monitored.

In the case where the investigator owns the data, Stanford will provide: 1) Guidance on assessing the risk profile of the dataset 2) a computational environment with security measures that are appropriate to the risk profile assigned to the datset 3) preset access permission frameworks for the risk profile assigned to the dataset 4) centralized access administration of the dataset.

In the case where Stanford is the data custodian and there are security or contractual reasons why Stanford or a researcher at Stanford is unable to share data, Stanford will make the code used to derive the cohort and the analysis available. In many cases, Stanford has the ability to expose meta-data or project construction and code without disclosing the actual data. If an investigator wishes to replicate the research or use the data in question, Stanford will work with the investigator to either secure the appropriate permissions so that they may use the data in the Stanford environment, or provide guidance on how to contact the data owner to access to the data.

Our hope is that this policy balances the need for free, discoverable and reasonable access to data, with proprietary and privacy considerations. We also hope to encourage researchers to post primary (rather than derivative datasets), and encourage transparency and replication of research.

* 1. Policies and procedures regarding the sharing, transmission, and distribution of PHS data files.

Proprietary data or data classified as moderate or high risk are shared within the confines of the PHS secure computational environment and data not transmitted outside of Stanford except under exceptional circumstances with the explicit written approval of the data owner or at the behest of the data owner. When access to data or analytic datasets is permitted by agreements with the data owners, data are only shared with individuals (including research collaborators) who have completed all the same security and training requirements as Stanford investigators as well as the requirements of their own university. Use of the PHS data by collaborators is governed by the same policies that apply to Stanford investigators. Requirements include: a) Completion of required Data Security, HIPAA and human subject protection (CITI) training(s) as applicable) and b) institutional review board (IRB) approval for their study or are included on an IRB approved study, c) demonstration of a “need to know” status with regard to the data and cannot practically work on the project without it and; d) signed data use agreement with Stanford PHS stating that they will only use the data for the stated research purposes and that they cannot download the data or share the data with any third party. Stanford’s policies governing the treatment and distribution of PHI and other high risk data can be found here: https://med.stanford.edu/irt/security/stanfordinfo.html

The data use agreement will stipulate that only output results from analyses can be downloaded – that is, the researcher agrees never to download datasets or subsets of those/analysis files and that all data outputs will conform to PHS’s (no cell smaller than 11) or data owner cell size policies, whichever is more restrictive. It is not permitted to share many PHS datasets outside of Stanford or Stanford Population Health Sciences. The term sheet for each dataset outlines these requirements and data access also requires all individuals accessing the data to have completed all requirements.

* 1. Data tracking system.

PHS data platform (powered by Redivis) has tools which track data cleaning and versioning in an automated fashion incorporated in this record are a) the database name, b) the date the encryption and cleaning processes were completed, c) the name of the encryption program utilized, d) their file sizes, and f) a description and code used for data cleaning and curation. Additionally, the Stanford PHS data platform automatically tracks user activity including a) datasets accessed, b) the date that the data were received, c) data reshaping activities d) who performed these activities e) files and variables accessed along with f) computational activity.

System and access logs are collected on each server and forwarded to a remote log server, per the requirements of Stanford’s Minimum Security standards for servers (see https://uit.stanford.edu/guide/securitystandards) using Splunk or equivalent service. Splunk is software that allows monitoring, searching and inspection of multiple system logs, across time; it is also a powerful tool for analyzing system logs to identify anomalies and trends. The PHS system administrators and the PHS Data Core have access to view and analyze log data associated with PHS servers, as does the university’s Information Security Office. Logs are retained for at least 18 months.

Data access will be monitored three ways depending on source and destination. File transfers and access via SFTP will be logged and sent to the Splunk service. File access on the storage system will be logged from the storage system and sent to the Splunk service. File access within the systems will also be monitored via the audited service and the Redivis tracking and auditing tools. PHS data do not leave the server environment. No investigator is permitted to download, print or otherwise remove data from servers which meet security standards as described above.

* 1. Policies and procedures for the physical removal, transport and transmission of data files.

In general, data are not physically removed from Stanford’s possession except at the behest of the data owner. In the case where Stanford will no longer host a dataset, the data will be purged from Stanford systems and the data owner will be provided with proof of data destruction as described in section 4.1.

In the rare event that data are physically transported, the original encrypted drives (if applicable) would be removed from our safe and shipped to the intended recipient in a manner meeting the security standards appropriate for that data. If the drives are not available, data would be transferred from our systems to an encrypted drive appropriate for the data risk category and shipped to the intended recipient in a manner which meets the security standards appropriate for the dataset.

* 1. Policies to tailor and restrict data access privileges based on an individual’s role on the research team.

The Data Core Manager will work with the PIs to determine which individuals on their research team require data access and which files they will require for their part of the proposed study. Individuals will be given access to the datasets and variables needed for the project. Individuals approved for access will have to complete the steps outlined in section 1.3. Access will be granted to individuals will only receive access to PHS data which are required for the approved analyses.

* 1. Technical safeguards for data access (including password protocols, log-on/log-off protocols, session time out protocols, and encryption for data in motion and data at rest).

As stated in sections 1.6, 1.7 and 1.9, the data from PHS (data) will be received by the Stanford PHS Data Core via secure FTP transfer or whatever method is preferred by the data owner and saved to a secure server that abides by Stanford’s computer and network usage policy (http://adminguide.stanford.edu/62.pdf) and information security policy (http://adminguide.stanford.edu/63.pdf), as well as minimum security standards for servers (https://itservices.stanford.edu/guide/securitystandards) for PHI and other sensitive data. The requirements associated with those standards include patching, server hosting in a data center, centralized logging, two-factor authentication, vulnerability scanning and mitigation, and intrusion detection. In addition, server system administrators are required to attend two days of information security training each year. The servers have also been outfitted with self-encrypting hard drives.

Computer encryption is conducted and verified using the Stanford Whole Disk Encryption (SWDE) service. The SWDE service is for both Windows and Macintosh computers that support native encryption. Once installed, all files are automatically encrypted. The data are protected while the computer is in standby or hibernation mode. This requirement applies to both Stanford and personally owned computers that are used for Stanford activities on the campus network. Computers on campus require SUNet authentication to enter. Personal computers require a 10 character PIN to open the computer. Additionally, all Stanford systems require two step SUNet authentication with a password. Passwords must comply with Stanford password guidelines (https://uit.stanford.edu/service/accounts/passwords). These standards also apply to all Stanford research collaborators.

Password Protocols:

Stanford recognizes that individual passwords often represent the weakest link to system access. Stanford’s password checking system is designed to address this; please see https://uit.stanford.edu/service/accounts/passwords for additional details. While Stanford implements a complex set of password rules, detailed at the previous URL, the institution also strongly recommends the use of pass phrases as an alternative. Beyond passwords, Stanford implements additional safeguards for systems for the storage and distribution of high risk data.

System logons will via SSH be authenticated via passwords, publickey and GSSAPI protocol combined with Duo two-factor authentication. Network firewall rules will require uses to be on the Stanford VPN and authenticated via SUNAC. The passwords and GSSAPI will be connected to Stanford’s Kerberos system and all Stanford SUNet password policies will apply. Every connection to the system will require a new Duo two factor prompt. If web applications are presented, the will be authenticated to via Stanford’s SAML2 instance, and Duo will be required.

SSH sessions will time out after 30 minutes of inactivity. To enable graphical interfaces, initially X2Go will be used through SSH for access to graphical environments. X2Go allows users to manager graphical sessions, reconnecting to existing sessions or starting a new session. This will permit longer term work in the graphical environment. To enable a better user experience and security, uses will be automatically place into a screen or tsmux session and reconnected when logging back onto the system. There will be a limit on ssh sessions per user; initially start at 5 sessions per user. The screen or tsmux session will lock the screen after a period of inactivity also.

Data transfers and access will be performed via SSL encryption, either via SSH or HTTPS. Data residing on hard drives will be encrypted via LUKS using hardware accelerated AES encryption.

Security Evaluations: Stanford PHS shall periodically (no less than annually) evaluate its processes and systems to ensure continued compliance with obligations imposed by Stanford policy, law, regulation, agreement or contract with respect to data confidentiality, integrity, availability, and security. Results of these evaluations will be documented and any remedial action indicated will be taken in a timely manner.

The Stanford Population Health Sciences Data Management Plan has been reviewed and approved by both the funding agency (Stanford PHS) as well as the Stanford Institutional Review Board (IRB). The data management plan will be reviewed no less than annually and updates will be made as indicated either by changes to the protocol or technological requirements for optimal data security.

File transfer activity is audited by the PHS Data Core on a monthly basis or as detection systems note unusual activity, whichever is more frequent. In the event of a suspected or reported file download, the researcher in question will be contacted. In the event any PHS (or other PHS) data has been downloaded to a computer or other device, the investigator will be required to purge the files from their computer. Disciplinary action up to and including termination of data access, employment or other sanctions will be taken as appropriate. All computers used to access PHS data must be encrypted and password protected as described above.

Both the regular and replica server adhere to all of the same security requirements. Stanford Nero servers are housed in a secure research data center, the Stanford Research Computing Facility, with keycard access limited only to authorized individuals. In addition, the server rack itself is locked; the key is only available to the system administrators of the server and the Data Center Manager. The Data Center has 28 cameras that capture motion 24x7. In case of facility power failure, supplemental power is provided by a generator. Testing of the ability to switch between commercial and generator power occurs monthly. The server has also been outfitted with self-encrypting hard drives. All statistical software necessary for analyses live on the server so that it is never necessary to download data onto local computers or other electronic devices. In addition to the above safeguards, server system administrators are required to attend two days of information security training each year and they only access the server from bastion host systems specifically secured by the Stanford Information Security Office. research collaborators with access to PHS data must meet or exceed Stanford standards for storage and computation for the risk profile assigned to the project data. Data will only be shared with collaborators when permitted by agreements (DUAs or Contracts) with data owners.

* 1. Research collaborators

In the event a Stanford investigator has a collaborator from another institution, access to data or analytic datasets is only permitted with individuals who have completed all the same security and training requirements as Stanford investigators as well as the requirements of their own university or institution. Use of the PHS data by collaborators is governed by the same policies that apply to Stanford investigators. Requirements include: a) Completion of required Data Security, HIPAA and human subject protection (CITI) training(s) as applicable and b) institutional review board (IRB) approval for their study or are included as personnel on an IRB approved study, c) demonstration of a “need to know” status with regard to the data and cannot practically work on the project without it and; d) signed data use agreement with Stanford PHS stating that they will only use the data for the stated research purposes and that they cannot download the data or share the data with any third party.

Investigators using PHS data are expected to make the final version of any resultant products of research (in most cases a peer-reviewed publication) as well as the code used to construct analytic datasets and analyses available to PHS. Insofar as it is possible, PHS makes code used for data management and analyses available to other PHS members and collaborators to increase efficiency and encourage replication. At the end of the project or proposed analyses, access to the project data will be terminated.

The signed data use agreement will stipulate that only output results from analyses can be downloaded – that is, the researcher agrees never to download datasets or subsets of those/analysis files or publish a formula which could be used to derive a cell with fewer than 11 individuals and that all data outputs will conform to Stanford or PHS cell size policies (no cell smaller than 11), whichever is more restrictive. All data will remain on servers which comply with or exceed Stanford’s standards for storage and computation for the assigned data risk category.

Procedures for Research Trainees must complete all the same security requirements as other PHS Data users at Stanford including encryption, signing a DUA, Data Security Training, and, where indicated, dataset specific training, CITI and HIPAA. All trainees must have an IRB approval or determination for the proposed project and an approved faculty PI. Additionally, trainees will receive extra instruction on data security and best practices as part of the classroom instruction. In general, trainees are initially granted access to lower risk datasets, either smaller (for example a 1 or 5% sample) or more redacted data to mitigate against risk. In the event a trainee project is of sufficient quality to publish or be included in further work, they are then granted access to the data as appropriate for their project under the same requirements as all other investigators. Trainees must be supervised by a Stanford faculty.

Computation will take place either on the PHS servers or a GCP environment approved for the risk profile assigned to the data. Activity will be logged and monitored. In cases where the method (eg, machine learning) or the study question (eg, rare diseases) require access to the full dataset, the trainee must complete a project form with an explanation for this requirement and, where applicable, a sample of their code. If a trainee has completed preliminary work of sufficient quality for a publication (eg, dissertation), then they will follow the same procedures as other PHS data users and submit preliminary results and a sample of their code. Applications for the full dataset will be reviewed on an ongoing basis by the PHS Data Core team for scientific merit, code quality and “need to know” thresholds. Assuming that the proposed project meets these criteria, full dataset access will be granted.

* 1. Procedures to house additional copy of the data including data transfer.

Data stored on Stanford servers are replicated nightly to a replica system in a different Stanford data center in Forsythe Hall on the Stanford campus. The server itself is housed in a secure research data center, with keycard access limited only to authorized individuals. In addition, the server rack itself is locked; the key is only available to the system administrators of the server and the Data Center Manager. The Data Center has 28 cameras that capture motion 24x7. In case of facility power failure, supplemental power is provided by a generator. Testing of the ability to switch between commercial and generator power occurs monthly.

1. **DATA REPORTING AND PUBLICATION**
   1. Notification of suspected incidents wherein the security and privacy of the PHS data may have been compromised. Includes policies and procedures for responding to potential breaches in the security and privacy of the PHS data.

In the event of a privacy incident the PHS Data Core will notify data owners within 36 hours Stanford PHS becomes aware that a determination that unauthorized access to, modification of, or disclosure of PHS Information and/or PHS Information Systems has occurred. A written resolution plan for any such incidents will be provided after any such incident. These plans will be drafted in collaboration with Stanford Research Computing, the Stanford Privacy Office and Stanford CTO.

In the event of an incident, a remediation plan will be drafted and executed according to Stanford PHS Data Management policy regarding management of PHS data: http://med.stanford.edu/phs/data-center-documents-.html.

* 1. Review and updates to data management plan.

Data Management plans are developed by the Stanford PHS Data Core leadership in partnership with the Director of the Stanford Center for Clinical Informatics and their staff. Data Management plans are then reviewed and by the Stanford IRB and the University Privacy Office.

Since the Data Management Plan is heavily influenced by data security technology and requirements, The Chief Technology Officer of Stanford Research Computing and the Data Core Manager will review the data management arrangements annually or whenever there is a significant change or improvement in data security or computational technology, whichever is more frequent. Any changes or updates to the Data Management Plan will also be included in the annual IRB renewal or modification, whichever is more frequent. If there is a change in Data Management Plan during the DUA period PHS will be informed within 5 business days of the proposed changes. Any changes requested by PHS will be incorporated into the DMP and submitted for approval at the time the DMP is submitted to the IRB and Stanford Privacy Office.

* 1. PHS cell suppression policy including a prohibition on publishing or presenting tables, outputs or formulas which display or back out to a cell with fewer than 11 individuals.

Stanford does not disclose direct findings, listings, or information derived from the file(s), with or without direct identifiers, if such findings, listings, or information can, by themselves or in combination with other data, be used to deduce an individual’s identity. Additionally, Stanford will not identify or report any identifiable pharmacy, provider, prescriber or health plan in any publication or present any cell or formula that can be used to back out a cell with fewer than 11 individuals. All data are reported in aggregate.

No cell (e.g. admittances, discharges, patients, services) 10 or less may be displayed or used in any publication. Also, no use of percentages or other mathematical formulas will be used if they result in the display of a cell fewer than 11. In the event that an investigator is unsure they meet the above criteria, they will submit written products for PHS review with the understanding that PHS agrees to make a determination about approval and to notify the user within 4 weeks after receipt of findings. PHS may withhold approval for publication only if it determines that the format in which data are presented may result in identification of individuals, providers or other cells smaller than 11.

1. **COMPLETION OF RESEARCH TASKS AND DATA DESTRUCTION**
   1. Process to complete the Certificate of Disposition form and policies and procedures to dispose of data files upon completion of its research.

If required by data owners, at the conclusion of the project, the research identifiable data will be purged from Stanford systems using packages like “shred”. Any retired data disks will be first “zeroed out” or reformatted. Any CD/DVD ROMs or other physical data storage devices will be destroyed or returned to PHS using PHS’s preferred delivery mode at that time. The Manager of Research IT will then complete the Certification of Disposition form and send it to the owner/vendor of the data as required. Stanford PHS Data Management policy regarding management of PHS data can be found here: http://med.stanford.edu/phs/data-center-documents-.html.

* 1. Policies and procedures used to protect PHS data files when individual staff members of research teams (as well as collaborating organizations) terminate their participation in research projects (which may include staff exit interviews and immediate access termination).

As outlined in section 1.4, the project PI will inform the PHS Data Core of any changes to personnel. In the event of an employee leaving of their own volition or being terminated prior to the completion of a project, the investigator or individual sponsoring the SUNet ID will terminate the sponsorship according to Stanford’s Network Access Control policy (https://uit.stanford.edu/service/sunac). For individuals who will no longer be employees, this termination is handled centrally in HR. Termination of SUNet sponsorship will prevent the employee or collaborator from having any further access to any Stanford system.

For non-employees or individuals remaining with Stanford but no longer working on a project which uses PHS data, the termination will be handled by the Stanford Data Core. The PHS Data Portal allows the Data Core Manager to assign permissions on an individual basis and terminate the ability for individuals to access PHS data and individual project folders on the Stanford PHS server. In special circumstances, a research collaborator may retain access to data for project completion or research collaborations. Such individuals must meet all the same training, encryption and security requirements as they did while Stanford affiliates including an addendum to the DUA or contract with the data owner where required.

As individuals are not permitted to download raw PHS data and access to data is restricted to those fields necessary for the analyses the individual was working on, termination of data access will prevent any further contact with PHS data.

Stanford PHS Data Management policy regarding management of PHS data and informing PHS of personnel changes can be found here: http://med.stanford.edu/phs/data-center-documents-.html

* 1. Policies and procedures used to inform data owners of project staffing changes, including when individual staff member’s participation in research projects is terminated, voluntarily or involuntarily.

Procedures for these cases are identical to those outlined above in Section 4.2.

* 1. Policies and procedures to ensure original data files are not used following the completion of the project.

Upon completion of the project, or completion of sub-analyses within a project, access to PHS data used for the project or analyses will be terminated. The Stanford data portal allows granular and configurable control of data access. As a second layer of protection, the University Network Access Control (SUNAC) service (<https://uit.stanford.edu/service/sunac>) allows Stanford to disable SUNet authentication which effectively removes all access. The service allows PHS Data Core staff (or the LNAs Local Network Administrator) to control remote access to departmental resources located behind University IT-managed firewalls.

In the event all data use agreements have expired and Stanford is required to destroy or return the data, the PHS Data Core will provide a certificate of destruction as described above.

1. **Management of Datasets with Identifiers (PHI and PII)**
   1. Trusted Third Party, linkage and deidentification of datasets which contain identifiers

Many research questions can only be answered by linking data by individual in order to assemble a complete picture of exposures, behaviors and outcomes. In the event that data linkage by individual will take place, The Stanford PHS Data Core acts as a Trusted Third Party to perform these linkages. Linkages that included identifiers will take place on a separate server or GCP instance and be conducted by a member of the PHS Data Core. Depending on the availability of identifiers in both data sources, direct or probabilistic linkage methods[[1]](#footnote-1) will be applied. Upon completion of the linkage, identifiers will be removed and the limited analytic dataset will be made available to the researchers through the usual channels.

The Stanford PHS Data Core team will have sole access to copies of the raw data and the algorithm used to link individual will be scrambled and encrypted and identifier that are stored in a both a restricted folder on the secure PHS server sequestered for this purpose and in hard copy in a locked safe as described in Section 1.9. The Stanford Data Manager screens all incoming linked data for identifiers and removes all but those necessary for research purposes, using the Safe Harbor Method or other accepted de-identification methods. The Safe Harbor method was laid out by the U.S Department of Health and Human Services in the context of HIPAA Privacy Rule and consists in removing or blurring data specific to 18 key identifiers[[2]](#footnote-2). Cleaned, de-identified files are then stored in files accessible for research. In some cases the PHS Data Core will also produce an anonymized version of the data. Data will be anonymized using accepted methods unclouding date jittering, 3 digit zip and similar. Depending on the source of the data, there may still be restrictions and requirements for fully anonymized data. The ability to access raw data which includes identifiers is tightly controlled and monitored by a small team at the PHS Data Core.

In the case where analyses require information which is technicially deemed identifiable, identifiers will be removed insofar as is possible. Datasets which contain identifiable information will be tagged RIF on the PHS data portal. RIFs or datasets which are rich enough that they are effectively identifiable will generally receive a high risk designation. Access permissions for these datasets will be set accordingly as described in Section 1.3.

As data are enriched, true de-identification becomes increasingly difficult and there is often an inverse relationship between anonymization and utility. Even if fully de-identified, datasets derived from PII or PHI often retain institutional or proprietary risk. Consequently, datasets derived from datasets containing PHI or PII will be evaluated individually and in many cases, treated as high risk, even if the risk of reidentifying an individual in the dataset is remote. Datasets which have been truly anonymized are often useful for feasibility or variable identification and for some datasets, we will make an anonymized version available. In all cases, the data owner will retain visibility into linkages and the ability to require additional protections or approvals.

1. Abramitzky, Ran et al. “Linking Individuals Across Historical Sources : a Fully Automated Approach.” (2018). [↑](#footnote-ref-1)
2. Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule, U.S Department of Health and Human Services,

   https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html [↑](#footnote-ref-2)