Guidelines for iPSYCH partition at the GenomeDK HPC cluster

Jakob Grove[1], Alfonso Buil Demur[2], Ditte Demontis[1], Thomas Als[1], Thomas Werge[2], Anders Børglum[1], Preben Bo Mortensen[3]

[1]Department of Biomedicine, Aarhus University

[2]Institute of Biological Psychiatry, Capital Region of Denmark

[3]National Centre for Register-based Research, Department of Economics, Aarhus BSS, Aarhus University

CONTENT

1. Introduction

This document describes how to access and use the GenomeDK Closed Zone Server (henceforth also referred to as *the server*) containing data from the Lundbeck Foundation Initiative for Integrative Psychiatric Research (iPSYCH) project and provides criteria and guidelines for working with microdata and data security.

The Danish population-based registers constitute an important research tool, enabling researchers to carry out representative population-based studies on, for example, the potential clustering of disease in families, and effect of parental income dynamics on health outcomes, and many more. Similar data are only currently available in Sweden, Finland and Taiwan, with varying differences in coverage periods and the extent of other available data can be linked to the population data. The Danish population-based registers will in connection with specific cohorts, intervention studies, biobanks and comprehensive genomics and

other biological data sets continue to provide the basis for significant knowledge relevant to the aetiological understanding and possible prevention of human diseases.

To access and successfully utilize these unique data-sources including person-level information on the iPSYCH population to the benefit of science and healthcare, is conditioned on compliance with Danish legislation, including complete confidentiality, while also ensuring flexible access necessary to conduct the required data analyses. Thus, the aim of this document is to ensure that personal data are used in compliance with legislation and the permission for granting access to data within iPSYCH, and within these constraints ensure a flexible access to these valuable data-sources.

The iPSYCH data is located at GenomeDK (the national high-performance computing facility for bioinformatics and life sciences managed by the AU/RM Genome Data Center and Center for Genomics and Personalized Medicine at Aarhus University) accessible through a NoMachine solution. Researchers access the data on the server through an encrypted connection with two factor identification from a personal computer. This is only possible from whitelisted IP numbers. Downloading person-level data (microdata) is neither possible, nor permitted, and thus all data processing can only be performed on the server. All users must read and comply with these guidelines. Users that do not comply with guidelines will have their access to data terminated.

Many datasets on the server contain microdata, i.e. individual level data on persons available from national registers or other resources. When conducting research on microdata, you must ensure that no microdata information is transferred to any unauthorized persons. Researchers may access data for the approved research only and must never reveal any microdata information to anyone outside the project. This is by far the most important criterion for any study of individual level data.

2. iPSYCH aim and permission

The Danish Scientific Ethics Committee, the Danish Health Data Authority, the Danish data protection agency and the Danish Neonatal Screening Biobank Steering Committee approved the iPSYCH study. This is in keeping with the strict ethical framework and the Danish legislation protecting the use of these samples. Permission has been granted to study genetic and environmental factors for the development and prognosis of mental disorders. Data may be used for research only and only within the boundaries of the research permission (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5754466/). The iPSYCH project is anchored at Aarhus University and the university is therefore responsible that iPSYCH comply with legislation and that all analyses are performed in compliance with the permission. In cases of doubt consult the Aarhus University representative. This is Professor Anders Børglum.

All personal data within iPSYCH accessed through GenomeDK have been deidentified, where the CPR-number is replaced by a scrambled personal identifier.

3. The server

GenomeDK is a high performance computer cluster designed for large scale genomics and bioinformatics research. iPSYCH has a secured partition with restricted access where the iPSYCH data is stored. Approved researchers can access the partitions through encrypted connection with two factor identification. Transfer of data out is blocked, and all data access is logged with username and time of reading the file.

The cluster is linux based, runs a modern queueing system providing access to ample computing power and very large and fast storage systems.

To access the iPSYCH partition at GenomeDK, researchers approved for access need to download and install a client for NoMachine and follow the steps described at http://ipsych.genome.au.dk/. This will set up a remote desktop accessible through a two-factor challenge.

4. Data Governance Experts

All projects and all new users working on the secure server will be assigned a Data Governance Expert. This person is appointed by the iPSYCH AU representative and supervises all issues related to accessing person-level information on the secure server. The Data Governance Expert adheres to the following criteria: a) Person with extensive experience in handling personal identifiable information at the server, and b) Employed at Aarhus University, and has at least 3 years of experience applying for and being granted approval to export statistically aggregated data from the server, and c) The user and the Data Governance Expert must collectively ensure that all analyses are performed in compliance with the permissions for the project and that they are necessary for the project.

When a new project is created, the iPSYCH AU representative appoints a Data Governance Expert. If possible, this is done among researchers already in the project. If not, a Data Governance Expert is selected to join the project. New users will be assigned Data Governance Experts based on the projects they join.

5. Microdata

Most data in iPSYCH project consists of microdata, which is personal, medical or biological data concerning individuals. All microdata must be treated as confidential information and must remain on the secure server. Even though all identifiers such as for example CPR numbers have been de-identified (replaced by scrambled identifiers), data is still microdata and may not be transferred out from the server. Even if you delete identifying variables such as the de-identified CPR number, it is still microdata and may not be transferred out from the server.

Researchers working on the iPSYCH project are obligated to treat all data as confidential information in accordance with the terms and conditions of the Danish Act on Processing of Personal Data. Confidential information is defined according to the General Data Protection Regulation, GDPR

(Persondataforordningen), the Danish Health Data Authority and Statistics Denmark's combined criteria. In general, it applies to any information that relates to less than 5 identifiable physical persons. This means that tables must contain at least 5 units per cell, and that all statistics must be based on groups of at least 5 individuals. In cases of doubt contact the Data Governance Expert.

On a case by case basis, it is possible to apply for permission to export certain types of aggregated data that on the face of it may not appear to abide by the above guidelines. Applications must be prepared together with the Data Governance Expert and sent to the iPSYCH AU representative. Examples could be principal component analysis plots of genotypes which are highly aggregated but reflect individual level information and certain rare genetic variant summary statistics. In such cases, care must be taken to ensure that it is not possible to identify individuals by combining with other information being exported.

6. Data security

Security rules

Only persons authorized by the iPSYCH AU's representative may access iPSYCH data.

New users must read this code of conduct for accessing and working with microdata and sign to testify that they have done so and will adhere to the described guidelines.

Microdata may not be sought extracted from the server in any way using whatever media. This also includes screen dumps, photographs, manual transcript of the screen, video, facetime, Skype, or any other method.

When connected to the server, the data on the screen may not be shown to persons who are not themselves granted access to the project.

When connected to the server, the computer shall not be passed on to unauthorized persons.

The passwords for accessing the server are strictly personal.

Lock the computer whenever you leave it – even if it is for just a minute. Make sure the local computer has auto lock, so that it locks automatically when not actively being used.

All analyses done with the data must be covered by the project description and must be necessary for the project.

Access to the server is possible only through the encrypted tunnel authorized through two factor challenge.

Access to data at the server is allowed only using a computer with up-to-date operating system and fully operational and updated antivirus software.

No attempts may be made to identify individual physical persons from the IPSYCH data.

Violating data security is a very serious breach of the agreement between the researcher and the iPSYCH AU representative. Non-compliance with the terms will exclude the researcher and potentially her/his host institution from access to iPSYCH data at the server temporarily or permanently.

Users must inform the iPSYCH AU representative and the Data Governance Expert immediately in case of any suspected breach or actual unauthorized use of sensitive data.

7. Transferring files

Getting results out from the server

Only aggregated results that do not allow individuals to be identified may be transferred from the secure server. It is of great importance that the researcher has made sure that files for export do not contain microdata information. All transferred files are stored for at least six months and are selected for inspection randomly or at the iPSYCH AU representative's discretion to make sure that they comply with the rules. If rules are violated, the penalty ranges from a personal warning to temporary loss of access for the individual and his/her host institution or even a permanent lock-out. For more details, please see below.

What can be transferred from the server?

- The governing principle is that it should not be possible to identify individuals. Hence, basing the aggregation on a minimum of 5 individuals is not necessarily sufficient. When transferring multiple outputs, make sure that it is NOT possible to identify individuals by combining two or more of the files that you are transferring.

In general files are permissible for export if all the included statistics are aggregated over at least 5 individuals, no count is less than or equal 4 (although 0 is OK), and it is not possible to identify individuals from the file alone or in combination with other files in the transfer.

Beware that exact medians, minimums, maximums or percentiles can identify individuals. Make sure to truncate values so they represent five (5) or more individuals. Also, be aware that it may be possible to identify individuals from outliers in a figure.

All output must be manually checked before transferred out. Transferring uncontrolled output is not allowed and considered a violation of the security rules. Users must know exactly what they are transferring and they are responsible for the content.

Experience from breaches of the security rules at Statistics Denmark has shown that they are almost always unintentional and very often because users were not aware that microdata was being written to the files exported. Therefore, special care must be given when exporting output files generated directly by software as opposed to manually curated files that allow a more direct control of the content.

Examples of information that CANNOT be transferred from the server?

- Output that in combination make it possible to identify individuals.
- Listing de-identified id numbers even if no other data is attached to it.

- Listing microdata even if it is not attached to any identifiers.

If you discover that the rules have been broken unintendedly or intendedly please contact the Data Governance Expert and the iPSYCH AU representative immediately. The immediate contact is important since it increases the possibility to contain the problem. It will be regarded as mitigating circumstances if the iPSYCH AU representative is informed about unintended mistakes by the perpetrator himself/herself.

If you are in doubt about the rules of transferring specific information from the server, then you should aggregate the output further or contact your project leader/Data Governance Expert. An unintended violation of the rules can have very serious consequences for you and the entire research environment.

Please also consult Statistics Denmark's paper for more information on Data Security https://www.dst.dk/da/TilSalg/Forskningsservice/Dataadgang and Statistics Denmark's guidelines regarding transfer of files from the servers https://www.dst.dk/da/TilSalg/Forskningsservice/hjemtagelse-af-analyseresultater. iPSYCH at GenomeDK adheres to Statistics Denmark's guidelines in cases of doubt.

Exporting files from the server

To export data from the GenomeDK server, first ensure that the data for export abide by the export rules, then follow this procedure:

1. Tar.gz the files to form a single compressed file.
2. Run the command gdk-export on the tar.gz file. This will send the file to a locked part of the sluicegate to wait for approval.
3. Send an email to cases@genome.au.dk, the Data Governance Expert and the iPSYCH AU representative Anders Børglum (anders@biomed.au.dk) with a brief description of what has been sent for export and stating that the files do not contain person identifiable information and is sufficiently aggregated.
4. Once export has been approved, the tar.gz file will be released from the gate and can be downloaded using sftp by logging in to 185.45.23.195. See https://genome.au.dk/zones/.

Example email:

I ask permission to export the following result file(s) [INSERT FILENAME(S) HERE] from the iPSYCH secure partition at GenomeDK. [INSERT BRIEF DESCRIPTION OF WHAT IS BEING EXPORTED AND THE PURPOSE OF DOING SO]

I am fully informed of the rules governing export of data from the iPSYCH Project and the Danish law on personal data "Persondataforordningen", and I confirm that the results do neither contain micro data nor individual level data.

A copy of all files sent for export is kept for at least 6 months at a location not accessible by the users, where they can be subject to inspection at random or at the iPSYCH AU representative's discretion. Inspection is carried out both manually and aided by algorithms.

Table 1. Examples of sanctions in case of violations of export rules

		Sanction against institution			**Sanction against researcher**	
Personal data transferred		**1st time for institution**	**2nd time for institution**	**Repeatedly for institution**	**1st time**	**Repeatedly**
a.	Technical accident, unintentional	1 month of quarantine for institution (there may be mitigating circumstances if, e.g., institution reported the mistake itself)	2 months of quarantine for institution	3 months of quarantine for institution	1 month of quarantine from all research projects (mitigating circumstances if, e.g., institution, reported the mistake itself)	3 months of quarantine from all research projects
b.	Deliberate action, to look at individual level data (e.g., in order to check for errors)	2 months of quarantine for institution	3 months of quarantine for institution	3 months of quarantine for institution	2 months of quarantine from all research projects	Permanent exclusion
c.	Deliberate attempt to identify	3 months of quarantine for institution	3 months of quarantine for institution	Permanent exclusion	Permanent exclusion	Not applicable

Sending files to the server

Files without microdata

To place information on the server that do not contain microdata, e.g. summary statistics, code or documentation, use sftp to copy the data to the sluicegate of the server at 185.45.23.195. Once that is done, the data can be found in the folder /data-lock/public/<username> and copied from there where it is needed. See https://genome.au.dk/zones/.

Files with microdata

Sending microdata to the server is done by data managers at the National Centre for Register-based Research on behalf of the iPSYCH AU representative. They will ensure that data are covered by the permission and that data the CPR-numbers are replaced by a scrambled personal identifier.