Operational guidance recommended for usage by researchers of online code repositories

Purpose of this guidance

This document sets out more detailed guidelines for researchers on the usage of online code repositories for UK Biobank approved research projects and should be read in conjunction with existing summary guidance at https://www.ukbiobank.ac.uk/media/1vhh2vip/github-3rd-party-sharing.pdf. These guidelines should be applied for researchers using any online code repositories, such as GitHub, BitBucket, SourceForge, or GitLab.

What are code repositories?

Online code repositories provide services which are commonly used by the research community to share and publish code: either within a specific group of users or more publicly, where access may be available to external users with or without accounts to the service in question.

The services, access to which these repositories facilitate, include:

- methods of version control for development of code;
- publication of other file formats, such as documentation and output of code;
- Access to view and download code and other related files.

Good practice guidelines

These guidelines set out good practice for researchers use of online code repositories such that researchers can ensure data security within their own organisation and research groups and also ensure that UK Biobank data is not inadvertently published or otherwise made available (outside of UK Biobank's normal access procedures).

- **Provision of roles**
  - Roles and permissions should be assigned such that reviews are required ahead of creating public uploads.
  - Write permissions (or higher) in repositories should only be provided to those who have undergone appropriate training in data security and standards.

- **Organisation of code and data**
  - To reduce the risk of inadvertent sharing of results, input data, or other private information on GitHub, a folder structure that separates code from input and output files is strongly recommended.
  - Other private information, such as IP addresses, passwords, or account credentials should also be kept outside of folders used for code. This reduces the risk of accidentally committing information which should not be publicly shared onto a Git repository.

- **Elements specific to repositories using Git**
  - .gitignore can be used to specifically instruct Git to not commit specific files or file patterns. Researchers can consider using a wide-ranging .gitignore, such as excluding all data file formats used for input and output (e.g. .csv, .txt, .vcf).
  - Git-secrets should be used prior to committing code to scan for passwords and credentials. Researchers can add further "secrets" to cover common patterns such as participant IDs, IP addresses, email addresses or usernames.

UK Biobank recommends the use of such practice by researchers generally and ***UK Biobank requires the adoption of such practices whenever UK Biobank data is involved.***