
DATA MANAGEMENT PLAN

Data sharing plan

Computing Environment: The UChicago computing environment provided by the Center for Research Informatics (CRI) will be sandboxed from the internet as well as other servers and data sources at UChicago. It will be accessible only to the PI, research assistant(s), software developer(s), and/or system administrator(s) who require access during the course of the project. **Box:** Research data will be stored and preserved for the duration of the grant using Box, which uses AES 256-bit encryption and is also FedRamp authorized and HIPAA compliant (<https://www.box.com/security>). Box provides file versioning, which helps mitigate issues such as file corruption. UChicago is committed to using Box as the institutional cloud storage tool. If the university were to switch cloud storage solutions, we will meet the security needs of this and all other grant work supported by Box.

Data and research resources generated in this project research will be made available to the research community, which includes both scientific and consumer advocacy communities, and to the public. This includes all data and research resources generated during the project's period of performance, including:

Unique Data, defined as data that cannot be readily replicated. For this project, examples of unique data include curated models of genomic change for different sub-types of Influenza A, for different geographical locations. **Final Research Data** defined as recorded factual material commonly accepted in the scientific community as necessary to document and support research findings. In our context, examples are sequence ids of strains we use for our modeling, and the particulars of validation experiments, including the metadata needed to replicate those experiments in the laboratory. **Research Resources** include, but are not limited to, the full range of tools that we would develop and use in the laboratory. In this project, such resources include all developed software for modeling and prediction.

We will deposit software in Github repositories, allowing easy installation of such software in compatible systems. We will also deposit models, metadata and software copy at Zenodo for long-term citable access to the research resources and products.

No data sharing agreement is required for this project, since the underlying data on which we will learn our models are publicly accessible with minor restrictions.

Complete enumeration of sequence ids as obtained from NCBI and GISAID will be submitted, which is sufficient to replicate the results if using our developed software. Also descriptions of inferred and curated models will be made available. Example software programs based on our open-source library will be provided as well.

No specialized file format is necessary for this project. All files will be shared as text files, csv files or compressed versions of those.

No specialized transformation is necessary.

The effort of the postdoctoral associate funded on this project will carry out the requirements of this plan, and his salary will be partially covered under the proposed budget.

