

SPIROMICS Policy on Datasets for Reproducing Manuscript Results

The Issue

An increasing number of journals are requiring authors of manuscripts to commit to making data available for other researchers to reproduce the results in the manuscript. A few journals even require data to be made freely available without restrictions. Some others require authors to choose between two statements with wording such as:

“The data, analytic methods, and study materials will be/have been made available to other researchers for purposes of reproducing the results or replicating the procedure.”

OR

“The data, analytic methods, study materials will not be made available to other researchers for purposes of reproducing the results or replicating the procedure.”

Although SPIROMICS has a substantial amount of data in repositories, such as BioLINCC and dbGaP, these repositories may not satisfy the needs for reproducing results in a specific manuscript because: (a) there is a lag between the data being available for use by SPIROMICS investigators and submission to the repositories; (b) because of consent restrictions and recoding of some variables to reduce the risk of identification of participants, the data in the repositories may not be identical to that used in a manuscript; and (c) details of analytic methods used in a manuscript, including computer code, are not included in the material submitted to the repositories.

SPIROMICS Policy

The SPIROMICS Genomics and Informatics Center (GIC), along with the Principal and Co-Investigators, are committed to efforts to enhance reproducibility of scientific research.

If the authors of an SPIROMICS manuscript are required or requested to provide a dataset and analytic methods for the purposes of reproducing results in the manuscript, the lead author should submit the necessary materials to the GIC for approval prior to submission to the journal. This should be done only when a request for a dataset is received however the GIC will always accept and archive datasets tied to a manuscript proposal and/or publication if desired. The GIC will ensure personal identifiers are removed and that proper consent is applied. To reduce the risk of deductive disclosure, the dataset submitted by the lead author should contain only the data that were actually used in the analyses for the manuscript.

When preparing a manuscript for submission to a journal such as PLOS One that requires data to be made available without restriction, consider using only participants who provided full consent (i.e. non-spiromics investigator use, commercial use, etc.). Once a dataset is submitted to PLOS One, SPIROMICS no longer has any control over how that dataset is used therefore this

upfront decision and restriction is imperative. For other journals, we will require a DMDA, and restrict use of the data to reproduce results in the manuscript as necessary.

The GIC's current subcontract with UCSF from NHLBI does not cover this work. If the number of requests for materials for reproducing results becomes substantial, the above policy may need to be re-visited and revised.

Documenting and archiving datasets and analytic methods

Even if a journal does not require datasets to be made available on request, it is good scientific practice that investigators working on a manuscript document their analyses, preferably including preparation of a statistical analysis plan before undertaking the analyses and details of exclusions applied to get to the subset of SPIROMICS participants to be used in the analyses. Further, once the work is complete the relevant analytic datasets and documentation should be archived securely. As noted above, sharing the relevant datasets and documentation with the GIC will always ensure documentation and archiving are handled consistently and appropriately.