

ARIC Policy on Datasets for Reproducing Results

The issue:

An increasing number of journals are requiring authors of manuscript to commit to making data available for other researchers to try 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 ARIC 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 ARIC manuscript because: (a) there is a lag between the data being available for use by ARIC 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.

ARIC Policy:

The ARIC Study investigators are committed to efforts to enhance reproducibility of scientific research.

If the authors of an ARIC 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 to the Coordinating Center the dataset, a corresponding data dictionary, and any information needed to understand how the data were used. If the journal requires computer code, this should also be submitted to the CC. Submission of material to the CC is needed only when a request for a dataset is received, either from the journal or from an individual after the journal has published the article with a statement that data may be obtained upon request. The CC will ensure personal identifiers are removed. To reduce the risk of deductive disclosure, the dataset submitted to the CC by the authors 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, use only data from those participants who provided full consent (that is, not “ARIC Only” or “Not for Profit”). Once a dataset is supplied to PLOS One,

ARIC no longer has any control over how it is used. For other journals we expect that we will be able to require a DMDA and so restrict use of the data to reproducing results in the manuscript.

(The CC's current contract 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.)

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 ARIC participants actually used in the analyses. Further, once the work is complete the relevant analytic datasets and documentation should be archived securely.