The goal of the NIMH Repository and Genomics Resource (NRGR) is to further the understanding of the genetic etiologies of mental disorders. The NRGR receives subject biological samples, such as blood, from NIMH-supported research projects, and processes these samples to DNA, RNA, cDNA, and/or cell lines, which can then be used for genomic analyses. The NRGR also receives, curates, and harmonizes clinical/phenotypic data for each subject. All data and biomaterials are made available to NIMH approved researchers through a secure web portal. This sharing of uniformly processed biological samples and curated clinical and genomic data from many cohorts leverages the NIMH investment in genetic studies and provides critical research power by making a very large body of data (>200K subjects) available for study of the genetic bases for individual mental disorders. The NRGR is supported by a $50M grant from NIMH to Rutgers.

Starting in March 2018, the clinical data curation functions of the NRGR will be transferred from a subcontract to Washington University to Rutgers University, under the direction of Dr. Linda Brzustowicz. A main goal of the new curation effort is to increase the use of automated data checking protocols. A flexible data checking program has been developed by our colleagues at the Information Sciences Institute at the University of Southern California. This "autoQC" program uses a web interface and allows researchers to check their data prior to submission. More specifically, the autoQC system expects data in pre-defined file formats, and will check the value of each entry against a "data dictionary" to make sure that the entry is the right type and within the right range. For example, the allowable values for sex may be "M" and "F". If anything else is found in the column of the data for sex, the autoQC program will generate an error.

Psychiatric genetics researchers use many diagnostic assessments in their studies. In order for this data to be checked by the autoQC system, we need data dictionaries for each assessment. The NIMH Data Archives (NDA), a clinical data repository, has developed data dictionaries for many of the diagnostic assessments we need to check. However, we have noticed that some of their dictionaries contain errors. For example, the allowable range of answers to a question may be 1-5, but the NDA dictionary may list 1-6. If there is an error in the dictionary, that will allow errors in the data to pass undetected, which is a problem for subsequent analyses.

We are looking for research students to help in our data curation efforts. Specifically, we need help with double checking the NDA data dictionaries by comparing the dictionaries to the documentation for the different clinical assessments. Students will also learn how to use the autoQC system and how to 'stress test' it to make sure it is functioning appropriately to catch different kinds of errors. These activities will be done in close coordination with the NRGR data curation team members in the Brzustowicz Lab. In addition, students will have the opportunity to participate in a supervised tutorial to learn the Python programming language and its application to the analysis of genomic data. While no programming skill is required for the initial data curation project, students who are interested and learn enough Python can participate in more advanced computational projects in future semesters. This is considered a 'dry' research project, as there is no laboratory bench work component. Students will learn important fundamental data management skills and have a chance to be part of an important team science endeavor supporting the national research program in psychiatric genetics. As much of the data to be checked pertains to psychiatric symptomatology, an interest in working with such data is a plus.