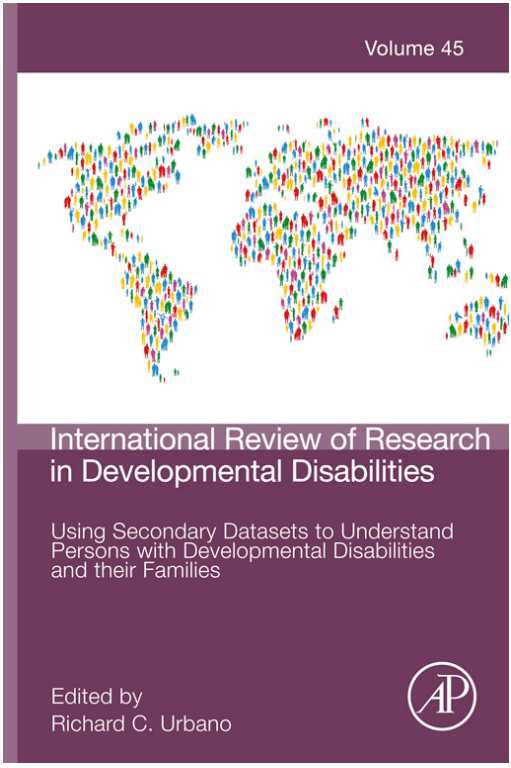

From: S.I. Novikova, D.M. Richman, K. Supekar, L. Barnard-Brak, D. Hall,
NDAR: A Model Federal System for Secondary Analysis in
Developmental Disabilities Research. In Richard C. Urbano, editor:
International Review of Research in Developmental Disabilities, Vol. 45,
Burlington: Academic Press, 2013, pp. 123-153.
ISBN: 978-0-12-407760-7

CHAPTER THREE

NDAR: A Model Federal System for Secondary Analysis in Developmental Disabilities Research

S.I. Novikova[*,1], **D.M. Richman**[†], **K. Supekar**[‡], **L. Barnard-Brak**[†], **D. Hall**[*]
[*]National Database for Autism Research, NIMH, OMNITEC Solutions, Inc., Rockville, Maryland, USA
[†]Texas Tech University, Lubbock, Texas, USA
[‡]Department of Psychiatry & Behavioral Sciences, Stanford University School of Medicine, Stanford, California, USA
[1]Corresponding author: e-mail address: svetlana.novikova@nih.gov

Contents

Abstract

The National Database for Autism Research (NDAR) is a human-subject data repository on tens of thousands of research participants. Approved researchers have access to an unprecedented volume of item-level clinical, genomic, and imaging data. Data are

International Review of Research in Developmental Disabilities, Volume 45
ISSN 2211-6095
http://dx.doi.org/10.1016/B978-0-12-407760-7.00003-7

shared quickly using both a common data standard and innovative tools for experiment definition, which provide the level of detail needed for efficient use of the repository. As described, early adopters have used it to conduct secondary data analysis. Now, with an ever-increasing volume of research data being made available, and new methods for data query, data download, and computation in place, this initiative is becoming vital to those interested in scientific discovery in autism or is being used as a model by other research communities.

1. INTRODUCTION

Widespread use of previously acquired research data for secondary analysis has provided enormous benefit across many scientific domains (Piwowar, Day, & Fridsma, 2007). In the health sciences, costly data-sharing efforts have often not met expectations for a variety of reasons, limiting investment in the infrastructure needed to support high-quality secondary analysis of research data (Poline et al., 2012). For developmental disabilities, data repositories generally focus on specific experimental techniques, such as genomics or imaging, or they are part of a clinical trial data collection protocol, which often limits the types of secondary analysis that can be performed. For complex disorders such as those related to developmental disabilities, a different data-sharing infrastructure is now emerging, one that provides broad support for data sharing across an entire research community. In autism, the National Database for Autism Research (NDAR) (ndar.nih.gov) is one of these models. This paper describes NDAR, how it has been used for secondary analysis to date, and how this resource can be best used in analyzing available data related to autism.

NDAR was established in 2006 by the National Institutes of Health (NIH) to provide a data-sharing infrastructure for the large Autism Centers of Excellence (ACE) program investment (NIH, 2007). For the ACE program, the NIH defined data-sharing schedules that each grantee agreed to meet. Generally, those terms define that all data, subject to appropriate informed consent, be submitted to NDAR and then shared. Data describing individual research subjects, such as raw and standardized scores for autism diagnostic (ADOS and ADI) and cognitive (intelligence quotient (IQ) and Vineland) measures, were expected to be submitted every six months throughout the course of each grant. In addition, raw data such as genomics sequencing, structural MRI, and EEG recordings were expected to be submitted. Submission of analyzed data was expected before

publication and would be shared with the research community when the data were published. These terms and conditions originally applied only to the ACE awardees, but have since been expanded to cover virtually all human-subject research related to autism (Sharing Data via the National Database for Autism Research, 2009). As a result of this data-sharing regimen, NDAR has quickly grown and now contains omics (e.g., genomics, metabolomics), neuroimaging, EEG, and clinical research data on 44,000 research participants across 400 measures. Approved researchers have unrestricted access to the data now shared in NDAR.

2. A COMPARISON OF NDAR WITH OTHER DATA REPOSITORIES

NDAR represents a new model for secondary analysis. One of its clear strengths is that it is a repository for data collected during studies funded by the NIH. The NIH has a very rigorous and well-designed peer-review process for selecting studies to be funded.

This approach provides a level of quality control with regard to participant recruitment, diagnostic verification, and procedures for collecting and analyzing data to answer the experimental questions proposed. High-quality NIH grant proposal reviews conducted by leading field scientists do not eliminate the possibility of unreliable or invalid data contained in NDAR. They do, however, increase the probability that the procedures implemented to collect and interpret data contained in NDAR are of sufficiently reliable quality to warrant inclusion in relatively large N studies where participants' raw data can be aggregated across multiple studies to answer questions that are quite different from the questions originally posed in the individual grants.

NDAR has several distinguishing characteristics that set it apart from other contemporary federally funded secondary data sets that sample populations with developmental disabilities, such as the Special Education Elementary Longitudinal Study (SEELS, 2005) and the National Longitudinal Transition Study-2 (NLTS-2, 2005). First and foremost, NDAR provides a dynamic rather than static secondary data set. Unlike the SEELS and the NLTS-2, data are continually contributed to NDAR and become accessible to authorized researchers weeks to months after they are received. SEELS and NLTS-2 data sets undergo months of vetting after data collection ends and before data are disseminated to authorized external researchers,

which can inhibit the progress of research. For instance, the first wave of SEELS data for the 2000–2001 academic year was not available for analysis to external researchers until 2005 (Godard et al., 2007). NLTS-2 may be considered as having a similar data turnaround time. Continually updated through the contributions of primary researchers, NDAR provides authorized researchers access to contemporary relevant raw data in a relatively short amount of time after its collection.

Second, NDAR provides data to the authorized researcher that have been contributed solely by other researchers. Data sets such as the SEELS and NLTS-2 are more multifaceted and provide information not only to researchers, but also to "parents, advocates, educators, researchers, and policy-makers" (Godard et al., 2007). So, while the current percentage of students with disabilities receiving free or a reduced-price lunch from their school may be of interest to policy-makers, this topic may not be of focal interest to researchers outside of the policy domain. Thus, NDAR is responsive to the needs of a research community rather than being subject to as much "political arithmetic" (Smith, 2008) as other publicly funded secondary data sources. While there is much promise in political arithmetic in addressing issues of social importance, the obvious pitfall has been that data analyses are not always meaningful beyond the original political intent (Smith, 2008).

Another strength is the amount of provided item-level data versus other federally funded data sets. For instance, the SEELS contains Scales of Independent Behavior-Revised (SIB-R, Bruininks, Woodcock, Weatherman, & Hill, 1996) data where only scale-level scores are provided. Thus, these data are not disaggregated to the item level for researchers to determine the psychometric sufficiency of these scales within certain populations of students, such as individuals with developmental disabilities. As a result, in using the SEELS, researchers may take a psychometric leap of faith that SIB-R scale scores are valid and reliable for their population of interest without access to item-level data. In contrast, NDAR contains a wealth of item-level data that permit researchers to examine the reliability and validity of a scale for their sample before committing to its use in their analysis for a population of interest. This access to item-level data appears to characterize the different audiences of these data sets well.

Given the relatively low incidence of autism and other developmental disabilities in the general population, the lack of probabilistic sampling methods, however, may be considered necessary to achieve appropriate sample sizes (Frederick, Barnard-Brak, & Sulak, 2012). The development of sampling techniques for rare or low-frequency populations continues

to address this issue of convenience in sampling individuals with developmental disabilities and other relatively rare populations (Lohr, 2010). While NDAR may not provide samples collected using complex survey design, NDAR data do appear to provide a greater depth of information of individual subjects as compared to other federally funded secondary data sets. Clinical background information with data such as IQ or ADOS scores for individuals with autism spectrum disorders would not be collected in other datasets due to the costly and labor-intensive nature of such assessments.

3. USING NDAR FOR SECONDARY ANALYSIS

For scientists interested in accessing data from NDAR, it is important to understand how the repository was designed and implemented. NDAR does not prescribe any research methods or the type of data to be collected. Instead, it only provides a platform for sharing the data that are being collected by studies funded by the NIH or any other funding agency (e.g., the State of New Jersey, Department of Defense, Autism Science Foundation). The only requirement is that the data being shared conform—or are harmonized—to the published data standard in autism (Hall, Huerta, McAuliffe, & Farber, 2012). This NDAR data standard, available on the NDAR website under the data dictionary, has supported data submission from over 100 projects and includes experiment definition in omics and other research areas.

The data submission approach is simple: when a new measure being collected in a research lab needs definition, NDAR works with the lab to create the definition, making the age of the participant and the subject identifier, called the NDAR GUID, part of the definition (Johnson et al., 2010).

Most of the data now being collected are defined in this way. However, while this approach for data sharing is efficient for data submission, it is far less efficient for data searches, especially across hundreds of individual measures. For example, a researcher interested in mining NDAR would need to query each type of data in NDAR and then combine the results to understand the types of data available for a particular subject. While it is possible to provide data by subject ID, it is cumbersome to query and extract data organized in this way. As shown, the researcher using NDAR would have to combine data across four separate instruments. While reasonable for four instruments, such an exercise becomes burdensome when interpreting the data across 400 instruments (Fig. 3.1).

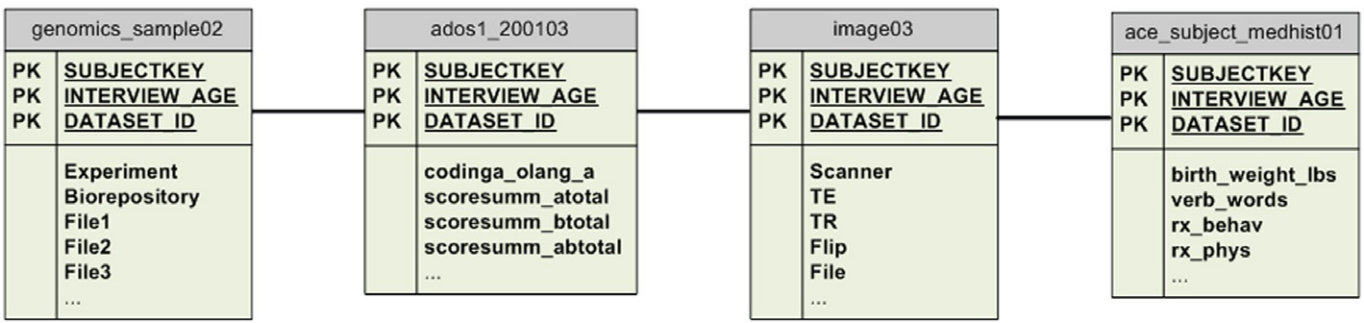

Figure 3.1 NDAR entity relationship diagram example.

Addressing this problem, NDAR implemented a series of database rules to collect data about a research subject and make that information available for query. The result is a transformed database model, called a star-schema, that provides efficient searches as the data in the fact table are derived from all of the data in NDAR and made available for filtering (Fig. 3.2).

Most of the data contained in the central fact table are actually derived summary elements, which allows both potential users and the general public to run sophisticated queries on the NDAR website. Fields such as clinical diagnosis and IQ that are embedded in a variety of measures can be extracted and made available for query. Thus, it is easy to determine what is available in NDAR for secondary analysis prior to applying for access. In addition, it allows data to be selected using these categories for detailed analysis by those having access.

Following are the steps used to determine what is available and how to extract data from the repository:

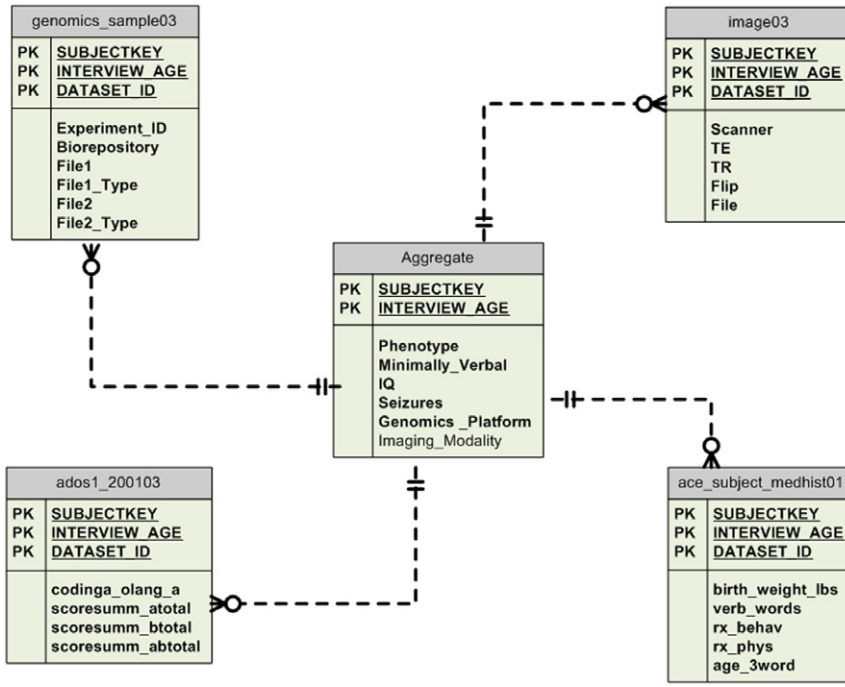

Figure 3.2 NDAR star-schema for query.

1. From the NDAR website at http://ndar.nih.gov, select the Query tab from the menu or go directly to http://ndar.nih.gov/query_data.html (Fig. 3.3).
2. Choose filtering options from the Select Data navigation (Fig. 3.4).
3. Choosing all in this case provides filtering options across a number of variables to identify cohorts (Fig. 3.5).
4. Filter your selection by phenotype or one of the other summary attributes that are presented. Selecting Severely Affected—a derived field used for initial filtering purposes only—displays the subjects available, reconstructs the pie charts to show numbers of subjects by age, and provides the option to download data. Many query options are provided and more will be added as the needs of the research community evolve. We welcome input on new query options (Fig. 3.6).
5. Select to download data. At this stage, only approved users may continue to have a package of data created for download (Fig. 3.7).

Also, additional attributes (e.g., diagnostic scores, IQ, and omic *p*-value) are provided. Given this approach, essentially any data contained in NDAR or one of NDAR's federated repositories can be selected for query. From this page, the Create Package button enables a user to select the cohort and download data. All data from all subjects selected will be provided. As a result, the package size of the data to be downloaded can be tens of terabytes, so opportunity to limit specific files or to move the data directly to a computational instance as explained later are becoming consistent.

Whether the data are downloaded locally to a researcher's computer or moved to a database in the Cloud, the data provided are the same as submitted to NDAR and shared. NDAR does not change any of the data it receives. If neuroimaging, omics, and diagnostic data for the subjects/cohort selected were requested, the data are provided in exactly the same way that they were received. However, other information about the subjects are provided as well, including any subject identifiers known to be used for the same subject, the fact table used to implement the star-schema model, and pedigree.

3.1. Beyond Secondary Analysis: Result Reporting

Making data available for secondary analysis is a primary objective of NDAR. However, when all data for a community are shared, an added and one could argue, essential—benefit to the system is the ability of the system to support data associated with published results.

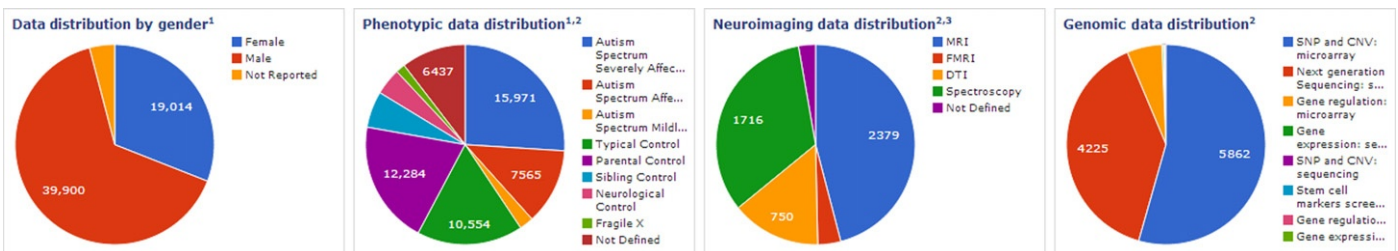

Figure 3.3 Categorical distribution of NDAR data.

Figure 3.4 Query filter navigation.

Often, publications offer insufficient detail on the data associated with the results. While many publishers have requirements that the data be shared so that any results can more easily be replicated, the reality is that data-supporting peer-reviewed publications are often not available for a wide variety of reasons (Savage & Vickers, 2009).

In autism research, the need to tie the publication back to the underlying data is particularly important for the replication of results. To address this issue, NDAR has developed the NDAR Study tool for the reporting of published results as well as results from unpublished hypotheses. The NDAR Study allows the author of a paper or a study to specifically define the individuals and the data from those individuals that were used. This simple definition allows others to easily reanalyze the data. When a paper with an NDAR Study is available in PubMed, the NDAR Study can be accessed directly from PubMed with one click (Fig. 3.8).

Soon we expect to assign a Digital Object Identifier for each study, which will allow another way for the data to be identified and potentially for data use to be tracked. A key advantage to the Study is that the subjects that were excluded from analysis are easy to identify when comparing all of the data available from a laboratory with the data used in the reported study. As the Study feature matures and the research community sees the value in following exactly what a laboratory has done during their analysis of the data, we believe that other data repositories will provide similar services. For those using NDAR for secondary data analysis, creating an NDAR Study is expected so that data provenance can be easily discovered.

3.2. Results from Omic Experiments

Computational approaches are becoming streamlined in autism omics and neuroimaging communities. While these processes are not yet mature, it is possible to organize analyzed data to better inform the research community. For neuroimaging, it is possible to compute across hundreds or thousands of subjects simultaneously using the resources available through the computational cloud, limited only by one's budget, and develop the methods to receive the analyzed data back into NDAR. As a result,

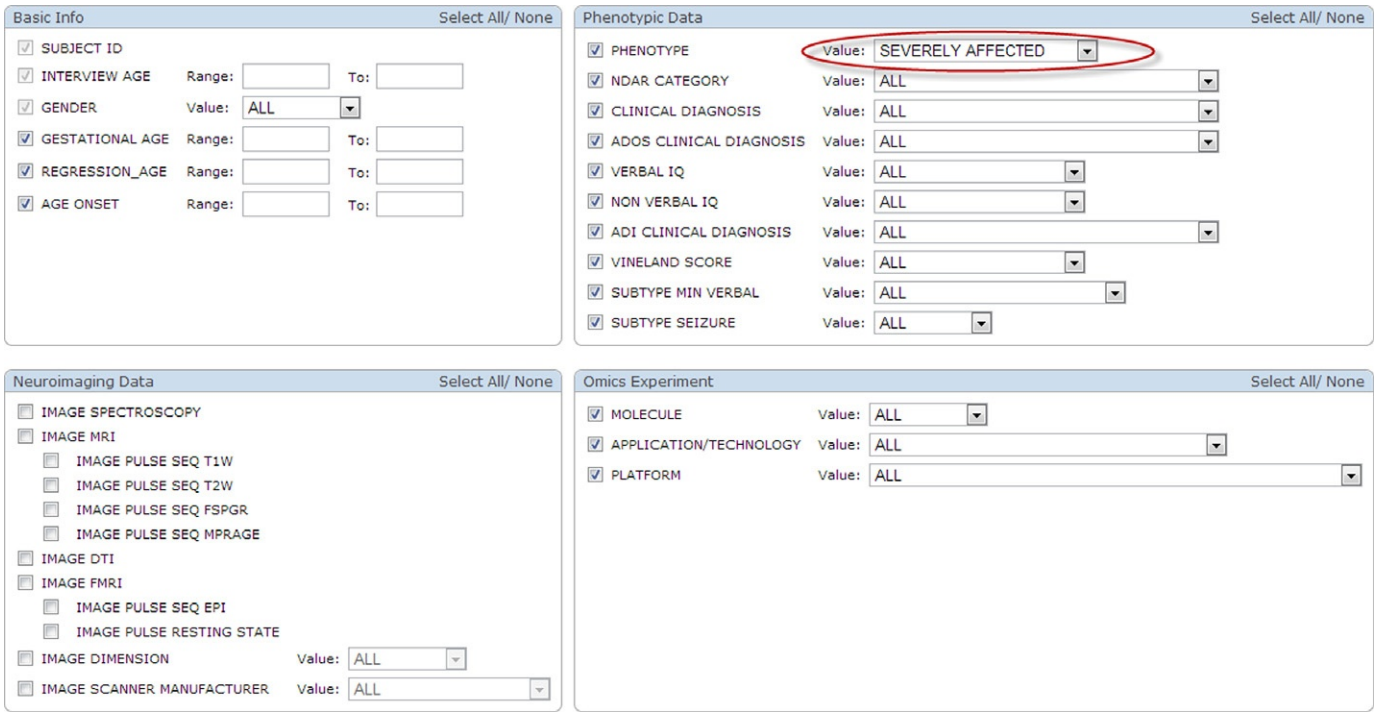

Figure 3.5 Query filter options.

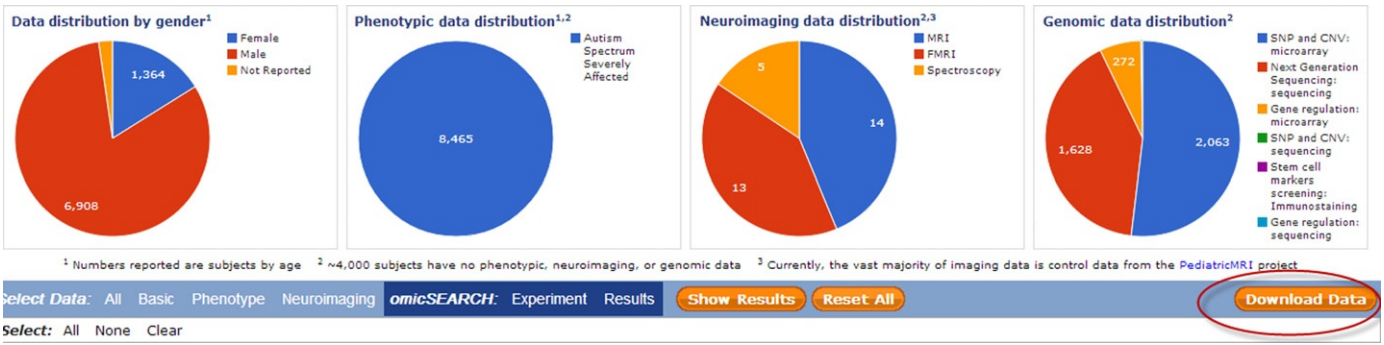

Figure 3.6 Download data interface.

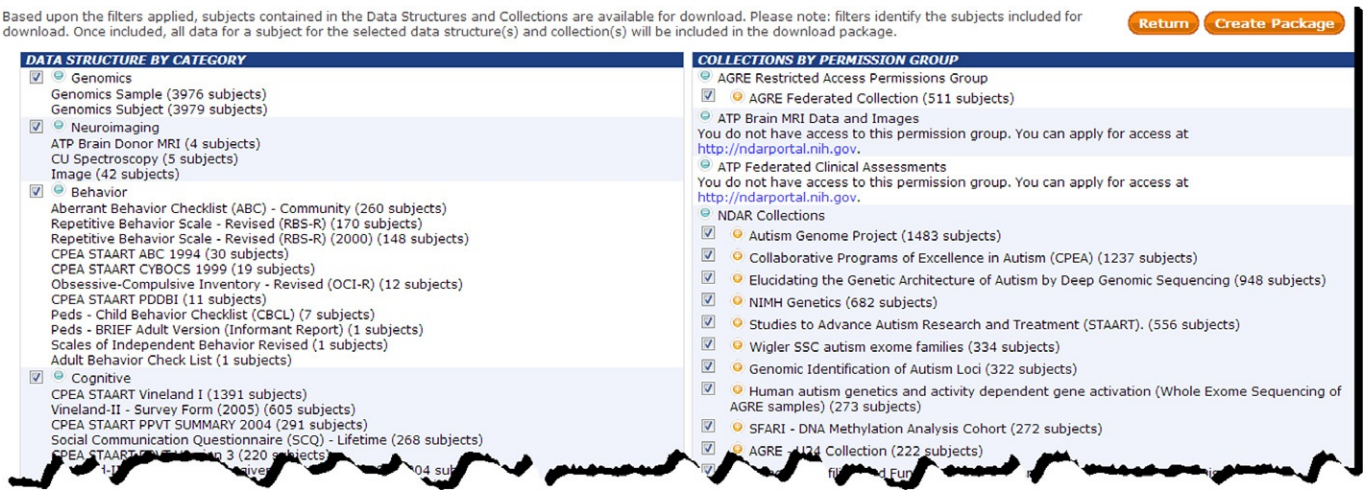

Figure 3.7 Package creation interface.

Figure 3.8 NDAR study listing.

volumetric measures, quality scores, IQ, and diagnostic scores can then be used for cohort selection. For neuroimaging, this is accomplished through results data structures that are now being developed to receive quality metrics from imaging pipelines. For omics, results have been standardized and curated through publications and provided through the omicSEARCH Results feature allowing filtering by name, location, affected region, and so on, as shown in Fig. 3.9.

As a result, the most significant alterations (e.g., snp, cnv, and gene) will become available for query. NDAR is now working on a system that will permit easy discovery of all of the alterations from a particular individual rather than just the newly discovered alterations that are reported in a paper.

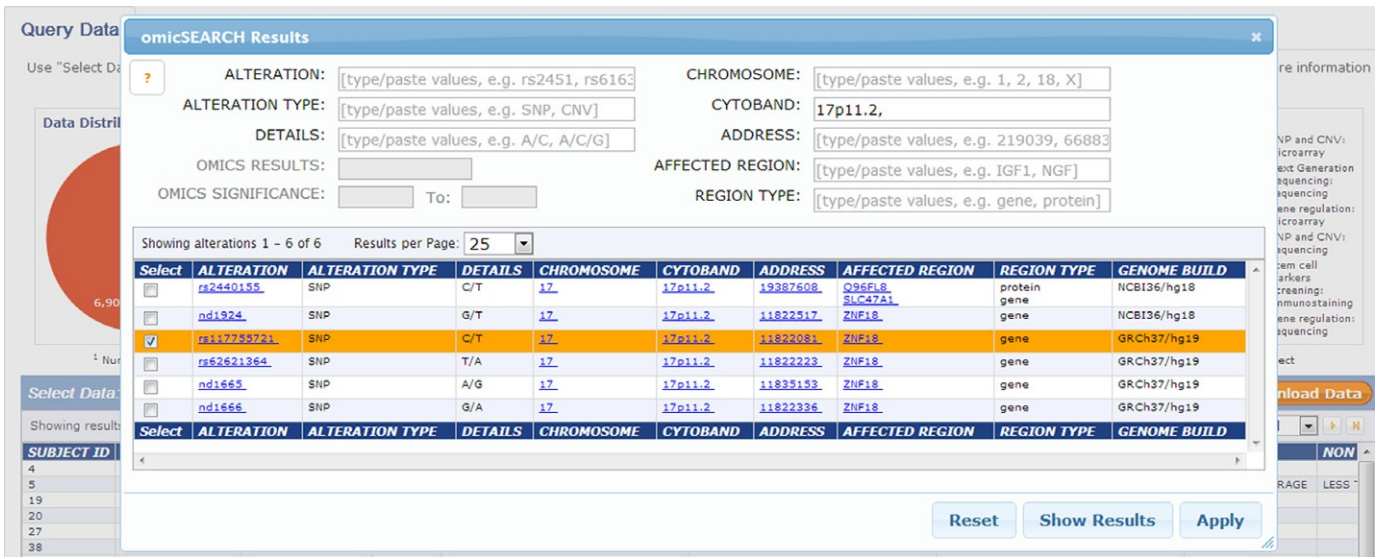

Figure 3.9 OmicSearch results.

3.3. Improvement Through Computational Integration

NDAR is a research repository containing an ever-increasing volume of data. Currently, NDAR holds over 100 Tb of raw data. Much more is expected given both the volume of data generated by the community and the need to begin capturing exposure related data.

Supporting this need, NDAR has moved its files (e.g., fastq, bam, dicom) into object-based storage (e.g., Amazon S3) and is working with neuroimaging and omics computational pipelines to make the most efficient use of these data by performing the secondary data analysis on NDAR data in the computational cloud. Clinical data contained in the NDAR database will remain at the NIH. Today, the files can be copied using the high-throughput capabilities of the Amazon cloud, but before long, references and permission to the files will be provided, eliminating the data-copying bottleneck. Using the parallel computational capabilities of cloud architectures, copy and process procedures can now be performed minimizing such bottlenecks.

Second, NDAR will continue to promote the implementation of computational pipelines used in omics and imaging analysis in autism where the data reside, encouraging scientists to compute on the data where it exists instead of downloading data to each laboratory many times, which is both time consuming and involves security concerns. Using available technology to save the operating system, computational software, and references to any of the data, it is possible and should one day be expected that the software and more importantly the configuration used to create published results will be saved and shared for result replication. Such capability is available within the Laboratory of Neuro Imaging pipeline today (http://www.loni.ucla.edu/) and should be a minimum expectation for any pipeline made available to others.

Lastly, the methods for data storage must be addressed. The cost of generating data from high-throughput experiments will continue to decrease, resulting in the need for more storage. As this trend continues, the return on investment of what data are saved and shared will gain increasing attention. To forestall this trend, NDAR is moving its big data to secondary storage (e.g., Amazon Glacier), significantly decreasing storage costs. It is likely that "all data" or even "all published data" cannot be stored indefinitely. However, using a centralized structure such as NDAR will enable the NIH, with input from the research community, to make informed decisions on data retention based on data quality and use.

4. NDAR MODEL AND LESSONS LEARNED FOR OTHER RESEARCH COMMUNITIES

NDAR has been in existence for six years and much of what is presented here was learned along the way. While not directly applicable to the researcher interested in using NDAR for secondary analysis, these points may provide insight on the data now contained in NDAR and help others looking to establish similar infrastructures.

4.1. Why Raw Data Are Necessary

The Human Genome Project showed the value of sharing data quickly, and other projects such as the Human Connectome Project (http://www.humanconnectome.org/) have adopted this "share widely before trying to publish" framework. In NDAR, data are made available at the earliest opportunity. Raw data being collected are shared throughout the life of the project. For efficiency reasons, NDAR chose to provide a submission timeframe of twice a year, but it is possible for labs to share sooner and more often. Clinical measures are shared early and often along with raw MRI, omics data, and initial EEG/eye tracking recordings. The sharing of raw data has two significant advantages. First, it allows secondary data analysis to start at the beginning. The other major reason to share raw data is to show the community which data were not used in analysis for a publication. For example, if a study includes 100 structural scans and only 60 are used in a publication analysis by the grantee, it is critically important to know why the other 40 scans were not used. For this reason alone, the sharing of raw data is essential.

4.2. Results Are the Most Important

In NDAR, it is now possible to link publications directly to the underlying data. NDAR has provided the infrastructure for this at very little burden on the investigator. Through variable definition and methods from the paper, and soon the saving of the software and system configuration, it is now possible to help eliminate ambiguity of methods from reported results. Both funding agencies and publishers should use the NDAR Study tool and others like it to promote data reanalysis.

4.3. Data Submission Is Different from Data Sharing

Often, data are withheld from submission indefinitely while a publication is undergoing review and then another publication comes along repeating the process. However, sharing the data with the research community is distinct from submission. Often, a publication will extend well beyond the end of the grant, and at that point there is no funding to support submission. Adapting to this part of the research process, data submissions schedules must be followed by every grantee with no exceptions. However, data sharing of results should happen at the time of publication, inherently requiring flexibility on when the data are shared with the community. NDAR and all other repositories understand this and will not share the data supporting a publication as long as it is accomplished within a reasonable amount of time.

4.4. Use the Common Definition

Whenever possible, projects should use the established data definitions made publically available on the NDAR website. Costs associated with data sharing are the greatest when projects define their own data standards, expecting others to then understand them. This is no longer an acceptable practice in autism research. For other research areas, the NIH now has a portal on common data elements that should be consulted prior to project inception (http://www.nlm.nih.gov/cde/). For many NIH funded projects, there is no good reason to use data collection instrument definitions that are different from those that are already available.

5. MAINTAIN DATA PROFESSIONALLY

Data, like the science used to create it, should be professionally acquired and managed. Relegating data sharing as a low priority and assigning the management of data to temporary lab support can have negative consequences.

5.1. Consider Using Existing Data and Computational Techniques in Research Aims

Data repositories are now available for secondary analysis and becoming easier to use. While in no way a substitute for *a priori* experimental analyses, review panels need to consider secondary analysis of data to increase sample size or corroborate possible findings (e.g., pilot data). Also, secondary analysis sources are ideal training ground for junior scientists.

5.2. All Data Can By Definition Be Shared

All data can be shared, including complex data types such as eye tracking, task-based fMRI, and EEG. No experiment is so complicated that the data cannot be made available to the research community. With proper planning, most experiments can be easily defined and shared without significant burden to the investigator. It is true that the initial definitions take some time, but the time spent is repaid at the end of the experiment when the data are being analyzed or when it is necessary to reanalyze data at a future date.

6. SECONDARY ANALYSIS RESULTS

NDAR is a data repository freely available to qualified researchers at institutions that have a Federal Wide Assurance for the protection of human subjects in place. To provide context to others on how it is being used for secondary analysis, two examples are offered. The first example comes from a group which not only uses NDAR as a source of data for secondary analysis, but contributes data to NDAR as well. The second group used NDAR as a source of clinical data and has published their results.

6.1. Secondary Analysis of Behavioral and Neuroimaging Data from NDAR

The main focus of the study by Supekar and colleagues (Supekar, Uddin, & Menon, 2013) was to characterize the behavioral profile and brain morphometry in girls and boys with ASD. ASD is diagnosed in females less often than males by a factor of one to four (Fombonne, 2003). Emerging behavioral accounts suggest that females are less severely affected than males on several measures of early social development (Lai et al., 2011). Remarkably, to date there have been no systematic attempts to characterize potential brain structural differences underlying the distinct behavioral profiles observed in females and males with ASD. Such work is critical for understanding the etiology of this heterogeneous disorder, as well as for understanding why the prevalence is lower in females. Supekar et al. address this critical gap by characterizing behavioral profiles and brain morphometry in girls and boys with ASD.

Behavioral and structural MRI data from 7- to 13-year-old girls with ASD and age- and IQ-matched boys with ASD were obtained from NDAR. As a first step, the investigators ran a query on behavioral data for children with high-functioning ASD, age between 7 and 13, and IQ

greater than 70. The query output was set to return age, gender, IQ, and phenotype along with scores on the Autism Diagnostic Interview, Revised (ADI-R). These query results yielded 632 children with ASD. Gender information was missing for 202 children and were therefore not included in the study. In the remaining 430 subjects, 72 were female and 358 were male. The mean age of girls with ASD was 9.1 years, and the mean age of boys with ASD was 9.1 years. This dataset comprised data from seven sites including Columbia University, New York University, the University of Pittsburgh, Stanford University, University of California—Los Angeles, University of North Carolina, and Yale University. The resultant aggregated dataset was exported to a csv file using NDAR's Download Manager. Next, the researchers obtained imaging data for this cohort by using a query by GUID. The query results yielded imaging data on 80 children, of which only 50 had high-quality imaging data. To ensure gender-distribution balance in the imaging sample as well as to increase the sample size, additional imaging data were acquired from the Autism Brain Imaging Data Exchange (ABIDE; http://fcon_1000.projects.nitrc.org/indi/abide). The NDAR imaging data along with the ABIDE imaging data were inputted to a customized subject matching algorithm (Uddin et al., 2013), which produced an age-matched balanced gender sample consisting of 25 female (mean age: 10.1 years) and 25 male children with ASD (mean age: 10.3 years).

Supekar and colleagues first examined the behavioral data on girls and boys with ASD. Comparing the cumulative ADI-R behavioral scores between the two groups, they found that girls and boys with ASD did not differ in overall severity of childhood autism. There were also no sex differences in social and communication deficits. However, females showed less severe restricted and repetitive behaviors (Supekar, Uddin, & Menon, 2013).

To delineate neural markers that underlie the unique behavioral profile in female children with ASD, the investigators next compared brain morphometry between girls and boys with ASD. Brain morphometry was assessed using the optimized voxel-based morphometry (VBM) method (Good et al., 2001) performed with the VBM5 toolbox (http://dbm.neuro.uni-jena.de/vbm). Prior to analyses, the structural images were resliced with trilinear interpolation to isotropic $1 \times 1 \times 1$ voxels and aligned to conventional AC–PC space, using manually identified landmarks, including the anterior commissure (AC), the posterior commissure (PC), and the mid-sagittal plane. The resliced images were spatially normalized to the Montreal Neurological Institute stereotactic space. Spatial transformation was nonlinear with warping regularization = 1, and warp frequency cutoff = 25. The

spatially normalized images were then segmented into gray matter (GM), white matter (WM), and cerebrospinal fluid compartments, with a modified mixture model cluster analysis technique (Good et al., 2001) with the following parameters: bias regularization = 0.0001, bias FWHM cutoff = 70 mm, sampling distance = 3, and HMRF weighting = 0.3. No tissue priors were used for segmentation. Voxel values were modulated by the Jacobian determinants derived from the spatial normalization such that areas that were expanded during warping were proportionally reduced in intensity. The investigators used modulation for nonlinear effects only (while the warping included both an affine and a nonlinear component). When using modulated images for performing subsequent group comparisons, the inference is made on measures of volume rather than tissue concentration (density). The use of modulation for nonlinear but not affine effects ensures that the statistical comparisons are made on relative (e.g., while controlling for overall brain size) rather than absolute volumes. The segmented (modulated) images for WM and GM were smoothed with an isotropic Gaussian kernel (10 mm full width at half maximum). The size of the kernel for smoothing was chosen as recommended by Gaser et al. for modulated images, since modulation introduces additional smoothing.

Discriminating brain morphometry patterns were identified using smoothed modulated GM images of girls and boys with ASD, using state-of-the-art multivariate pattern analysis (MVPA) (De Martino, Valente, et al., 2008; Uddin et al., 2011). To identify discriminating brain regions, the investigators used MVPA instead of conventional univariate analyses for the following reasons. First, the MVPA framework allows examination of datasets comprising data acquired at different locations at different scanners with different acquisition protocols. A traditional univariate analysis, with its underlying statistical assumptions of statistical independence, does not afford analysis of multisite data. Critically, MVPA analysis is very well suited for analysis of publicly available data from data repositories such as NDAR, which comprise data collected at multiple sites. Second, the MVPA technique provides greater sensitivity than the univariate VBM approach, as it evaluates spatial patterns in multiple voxels at a time. Specifically, a multivariate analysis that takes into account spatial patterns in the data would detect differences here, while the univariate would fail. Thus, the improved sensitivity is due to the consideration of spatial patterns of group differences, above and beyond those detectable at the individual voxel level.

The investigators performed the MVPA analysis using LIBSVM software (Chang and Lin, 2011). Inputs into the MVPA were the smoothed GM

maps computed from the VBM analyses. The MVPA method uses a nonlinear classifier based on support-vector machine algorithms with radial basis function (RBF) kernels. Briefly, at each voxel (vi), a $3 \times 3 \times 3$ neighborhood (searchlight) centered at vi was defined. The spatial pattern of voxels in this block was defined by a 27-dimensional vector. For the nonlinear SVM classifier, two parameters were specified, C (regularization) and α (parameter for RBF kernel), at each searchlight position. The investigators estimated optimal values of C and α and the generalizability of the classifier at each searchlight position by using a combination of grid search and cross-validation procedures. In earlier approaches, linear SVM was used and the free parameter, C, was arbitrarily set. In the current work, however, free parameters (C and α) were optimized based on the data, thereby designing an optimal classifier. In the M-fold (here $M=10$) cross-validation procedure, the data were randomly divided into M folds. $M-1$ folds were used for training the classifier and the remaining fold was used for testing. This procedure was repeated M times wherein a different fold was left out for testing each time. Class labels of the test data were estimated at each fold and average classification accuracy was computed for each fold, termed cross-validation accuracy (CA). The optimal parameters were found by grid-searching the parameter space and selecting the pair of values (C, α) at which the M-fold CA was maximum. To search for a wide range of values, the investigators varied the values of C and α from 0.125 to 32 in steps of 2 (0.125, 0.25, 0.5, 2, 16, 32). The resulting 3D map of CA at every voxel was used to detect brain regions that discriminated between the two participant groups. Under the null hypothesis that there is no difference between the two groups, the CAs were assumed to follow the binomial distribution $\mathrm{Bi}(N,p)$.

Using MVPA analysis, they found that GM in several cortical regions could discriminate girls and boys with ASD (Supekar et al., 2013). Additionally, they found that the GM volume in a subset of these regions was correlated with scores on the Repetitive/Restrictive Domain of the ADI-R such that the girls with the least impairment in the Repetitive/Restrictive domain showed greatest GM volume in regions involved in motor function (Supekar et al., 2013). No such relationship was observed in boys.

Supekar et al. claim to have discovered evidence for distinct behavioral profiles in girls with ASD, compared with boys. In addition, they also show a link between these behavioral differences and brain structural differences, demonstrating for the first time that at earlier ages closer to disorder onset, the brain in female children with ASD is structured in ways that may

contribute to reduced behavioral impairments. This study was performed with data available from the NDAR database, rather than with data that were specifically acquired to explore this research question. This case serves as a prime example of how such secondary data analysis can propel scientific discovery and demonstrates the high quality of data currently in NDAR.

6.2. Secondary Data Analysis Clinical Phenotype: Predictors of Self-Injury

The purpose of Richman et al. (2013) was to replicate and extend previous research on risk factors associated with self-injurious behavior (SIB) exhibited by individuals with ASD using a large and diverse sample. Specifically, items from the ADOS, several standardized measures of intellectual abilities, and the Aberrant Behavior Checklist (ABC) were used to analyze the following variables as predictors of SIB: (1) degree of intellectual impairment, (2) severity of autism symptomatology, (3) impulsivity, (4) hyperactivity, (5) stereotypy, and (6) low affect.

The series of recent articles on the relation between impulsivity and SIB exhibited by individuals with ASD and other developmental disorders (Arron, Oliver, Moss, Berg, & Burbidge, 2011; Richards, Oliver, Nelson, & Moss, 2012) were very intriguing because they incorporated relatively novel variables in addition to the traditional variables (e.g., level of intellectual impairment, communication impairment, and presence of specific genetic disorders) that are highly correlated with the increased probability of SIB in individuals with intellectual and developmental disabilities (IDD). NDAR provided access to a large enough sample of individuals with ASD to conduct the appropriate data analysis for predictor variables for SIB. The group was able to identify 617 individuals from 45 studies with 22,466 participants with a confirmed ASD diagnosis and a standardized measure of a wide variety of variables that may predict SIB in ASD (Richman et al., 2013).

6.2.1 Sample and Measures

A sample of 617 individuals with ASD diagnoses was derived from the following collection IDs (along with submitters): NDARCOL0000011 (University of Washington ACE), NDARCOL0001854 (Studies to Advance Autism Research and Treatment (STAART)), and NDARCOL0000001 (University of Illinois at Chicago ACE: Translational Studies of Insistence on Sameness in Autism). The process for finding the 617 participants consisted of several "queries" to NDAR to isolate participants that had ADOS and ABC data, along with any standardized measure of intellectual abilities.

Latent constructs were created from raw data consisting of individual items of the ABC community version (Marshburn & Aman, 1992). The ABC has been used as a measure of problem behavior severity in individuals with IDD (Lopez, Lincoln, Ozonoff, & Lai, 2005). Items on the scale describe maladaptive behaviors and caregivers rate the behavior on a four-point scale. Initial factor analyses of the items on the checklist yielded five subscales: irritability, lethargy, stereotypic behavior, and hyperactivity (Aman, Singh, Stewart, & Field, 1985a). Previous research has shown that the ABC has sufficient psychometric properties (Aman, Singh, Stewart, & Field, 1985b), and validity of the subscales has been verified in a sample of individuals with ASD (Brinkley et al., 2007). SIB scores were obtained by constructing a factor of three items on the ABC as validated by Brinkley et al. (2007) with a sample of individuals with ASD. For the purpose of Richman et al. (2013), ABC items comprising the factor for stereotypy were the same as the ABC subscale stereotypy. For the factors of hyperactivity and impulsivity, ABC items for the hyperactivity subscale were utilized to create the two latent variables. The low-affect factor was developed from items on the irritability subscale of the ABC.

Symptom severity of ASD was measured using scores from the ADOS (Lord et al., 2000). Although the scores were not originally normalized to use as a measure of relative symptom severity, previous research has shown that higher scores suggest a greater number, and severity, of core ASD symptomatology (Gotham, Pickles, & Lord, 2009).

Intellectual abilities, measured by IQ scores, had a mean of 80.98 (SD = 28.46) with values ranging from 36 to 143. IQ scores were derived from six different measures of IQ employed in the three different NDAR collections/studies. While NDAR presents the opportunity for researchers to examine research questions with large sample sizes, the use of six different measures of IQ (e.g., WISC, Stanford Binet) does introduce a degree of measurement error, which should be acknowledged as possibly influencing the results of any study, especially in cross -validation.

6.2.2 Data Analysis

Structural equation modeling (SEM) techniques were used in testing our model of SIB via M*plus* (Muthén and Muthén, 2008). SEM consists of a set of multivariate techniques that are confirmatory rather than exploratory in testing whether models fit data (Byrne, 2011). SEM has three major advantages over traditional multivariate techniques: (1) explicit assessment

of measurement error; (2) estimation of latent (unobserved) variables via observed variables; and (3) model testing where a structure can be imposed and assessed as to fit of the data. Most multivariate techniques inadvertently ignore measurement error by not modeling it explicitly, whereas SEM models estimate these error variance parameters for both independent and dependent variables (Byrne, 2011). In addition, SEM permits the estimation of latent variables from observed variables; thu, the creation of composites takes into account measurement error. Finally, fully developed models can be tested against the data using SEM as a conceptual or theoretical structure or model and can be evaluated for fit of the sample data. As an advanced statistical technique, SEM requires sample sizes of at least 200 to examine basic models. More complex models would require even larger samples in order to achieve statistical power. Thus, NDAR provides researchers the opportunity to utilize these sophisticated statistical techniques through the availability of larger sample sizes.

In SEM, models are first evaluated for fit. Upon satisfying fit, individual paths may be evaluated. Four statistics were considered to evaluate model fit. Comparative fit index (CFI) and Tucker Lewis index (TLI; also known as the non-normed fit index) values were evaluated, with values of 0.90 and above indicating an acceptable level of model fit (Weston and Gore, 2006). Finally, the standardized root mean residual value was considered in evaluating model fit such that values of 0.08 or less were considered indicative of acceptable model fit (Hu and Bentler, 1999; Schermelleh-Engel, Moosbrugger, & Müller, 2003).

Missing data were handled using full-information maximum likelihood (FIML) as the method of estimation in testing the model. FIML does not provide an imputation of missing data values, but rather estimates coverage of missing data at the covariance matrix level (Allison, 2003). As an extension of maximum likelihood, FIML uses all possible data points during data analyses. FIML is also known as raw data maximum likelihood, where observations are classified into different missing data patterns, "with all patterns subsequently being analyzed into a multiple group design with appropriate constraints across the groups," (Stoel, van den Wittenboer, & Hox, 2003). Enders and Bandalos (2001) have indicated that FIML is superior to listwise, pairwise, and similar response pattern imputations in handling missing data that may be considered ignorable. Multiple imputation methods were utilized in estimating missing data for the variable of intelligence. Missing data techniques could not be utilized without the availability of larger sample sizes such as through NDAR.

6.2.3 Results

First, conducting confirmatory factor analyses for each construct assessed model fit. While previous literature indicated the psychometric sufficiency of ABC items, there has been limited research on the psychometric properties of the ABC when applied to samples consisting of individuals with ASD (Brinkley et al., 2007). As a result, we assessed each construct individually to confirm the measurement model. Values for Cronbach's alphas ranged from $\alpha = 0.80$ to 0.94, revealing an acceptable internal consistency of scores to provide evidence for reliability.

CFI and TLI values were 0.92 and 0.93, respectively, indicating acceptable model fit of the data. In addition, the value for the SRMR was 0.05, indicating acceptable model fit as well. Thus, it appeared that the model fit the data sufficiently well to proceed with evaluating the predictor paths.

The figure below contains the standardized path values and the associated levels of statistical significance. As represented in the figure, we statistically controlled for intelligence and severity of ASD symptoms with IQ and ADOS scores, respectively. IQ scores were significantly negatively associated with SIB with a standardized path coefficient value of −0.39. This value indicates that as IQ scores increased, SIB decreased and vice versa. ADOS scores were not significantly associated with SIB, but ADOS and IQ scores were significantly associated with a standardized path value of 0.26. It should be noted that ADOS scores did not predict IQ scores, nor did IQ scores predict ADOS scores when testing competing models.

In evaluating predictors of SIB, it should be noted that two statistically significant predictors were revealed: impulsivity and stereotypy. Impulsivity was positively associated with SIB with a standardized path value of 0.46. This path value indicates that as impulsivity increases, SIB increases. Additionally, stereotypy was positively associated with SIB with a standardized path value of 0.23. This path value indicates that an increase in stereotypy is also associated with an increase in SIB. Among the predictors that were modeled, impulsivity, followed by stereotypy, appeared to be the most salient variables associated with SIB among individuals with ASD after IQ and severity of autism symptoms were controlled for.

6.2.4 Summary

Richman et al. (2013) extracted archival data from NDAR and replicated two primary findings from a series of recent studies conducted by Chris Oliver and colleagues (Richards et al., 2012). That is, severity of impulsivity and

stereotypy both independently predicted severity of SIB. Increased SIB was associated with greater impulsivity and more chronic stereotypy. It is important to recall that these data reflect caregiver reports of responding, not direct measurement of responding. However, the secondary data analysis provides a strong foundation for further investigation into impulsivity, stereotypy, and SIB, using direct methods of measurement.

As noted by Richards et al. (2012), the consistent finding of impulsivity as a risk factor for SIB suggests that some cases of SIB may be exacerbated by motor impulse control difficulties, or impulsive decision making, commonly associated with limitations in executive functioning often observed in individuals with attention deficit hyperactivity disorder (Barkley, 1997; Lambek et al., 2010). As discussed by Turner (1999) and Oliver, Petty, Ruddick, and Bacarese-Hamilton (2011), executive dysfunction may lead to increased difficulty inhibiting responses. Thus, it makes intuitive sense that some cases of SIB in ASD can be reduced if impulsivity can be reduced through pharmacological or behavioral approaches, or a combination of the two approaches. This is merely a hypothesis at this time, but it certainly warrants additional research to confirm or refute the hypothesis. We believe that this type of finding from secondary data analyses from NDAR data will eventually accelerate the speed of scientific discoveries for clinical behavioral phenotypes because hypotheses can be tested with preexisting samples of individuals with confirmed ASD diagnoses. This is a distinct advantage because it will allow scientists to move from hypothesis development, to testing the potential validity of hypotheses with "data in hand" from NDAR, and then moving on to more advanced hypotheses to be tested in future studies and grant proposals to federal and private foundations (Fig. 3.10).

7. CONCLUSION

As shown, early adopters of NDAR have used it to advance scientific discovery in autism. However, with new techniques for query, data download, and computation, the use of this repository is positioned to be an integral part of the research lifecycle in autism. Furthermore, the model presented here can be replicated to other research communities, where greater numbers of participants, computational techniques, and replication of results are important.

NDAR provides an infrastructure for harmonized data sharing containing an unprecedented amount of research data in autism. The depth

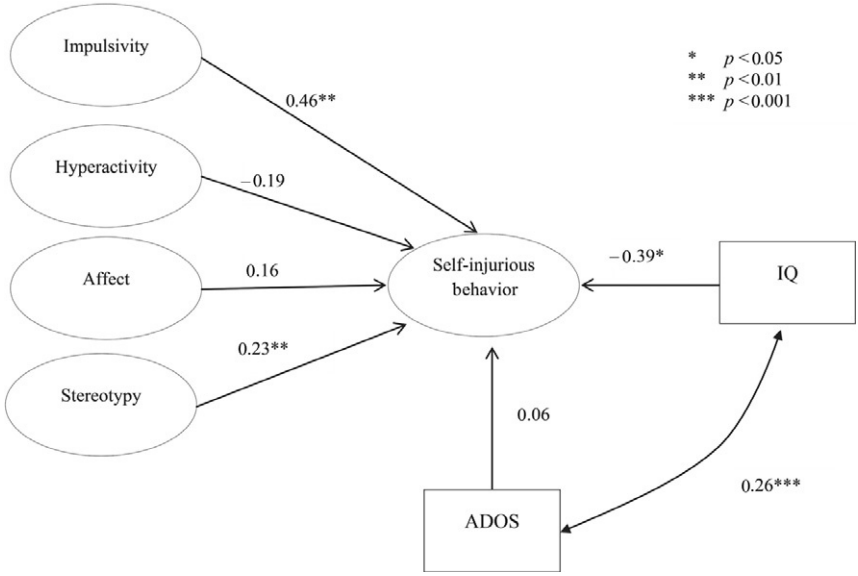

Figure 3.10 Path diagram for self-injurious behavior. *The figure depicting predictors of SIB and selected portions of this manuscript describing predictors of SIB were previously published in Richman et al. (2013). The figure and selected sections of Richman et al. (2013) were reprinted with permission from Blackwell Publishing Ltd. and John Wiley & Sons Ltd., UK.*

and breadth of data available will continue to grow offering greater opportunities for those interested in using it for secondary analysis. For better or worse, the data available directly represent the data being acquired in autism research, but data quality continues to improve as the community emphasizes data sharing for the research data now being collected. Improvement in the quality of rich datasets can be automated through available computational pipelines, which should have a profound impact on secondary analysis including any results from these techniques. As this trend continues, repositories such as NDAR are positioned to become central for use in secondary data analysis and the reporting of research results.

REFERENCES

Allison, P. D. (2003). Missing data techniques for structural equation modeling. *Journal of Abnormal Psychology*, *112*(4), 545–557.

Aman, M. G., Singh, N. N., Stewart, A. W., & Field, C. J. (1985a). The aberrant behavior checklist: A behavior rating scale of treatment effects. *American Journal of Mental Deficiency*, *89*, 485–491.

Aman, M. G., Singh, N. N., Stewart, A. W., & Field, C. J. (1985b). Psychometric characteristics of the Aberrant Behavior Checklist. *American Journal of Mental Deficiency*, *89*(3), 492–502.

Arron, K., Oliver, C., Moss, J., Berg, K., & Burbidge, C. (2011). The prevalence and phenomenology of self-injurious and aggressive behaviour in genetic syndromes. *Journal of Intellectual Disability Research*, *55*, 109–120.

Autism Centers of Excellence (ACE) program (http://www.nimh.nih.gov/science-news/2007/nih-funds-new-program-to-investigate-causes-and-treatment-of-autism.shtml.

Barkley, R. A. (1997). Behavioral inhibition, sustained attention, and executive functions: Constructing a unifying theory of ADHD. *Psychological Bulletin*, *121*(1), 65–94.

Brinkley, J., Nations, L., Abramson, R. K., Hall, A., Wright, H. H., Gabriels, R., et al. (2007). Factor analysis of the Aberrant Behavior Checklist in individuals with autism spectrum disorders. *Journal of Autism & Developmental Disorders*, *37*, 1949–1959.

Bruininks, R. H., Woodcock, R. W., Weatherman, R. F., & Hill, B. K. (1996). *Scales of independent behavior-revised*. Itasca, IL: Riverside Publishing.

Byrne, B. M. (2011). *Structural equation modeling with MPlus: Basic concepts, applications, and programming*. New York, NY: Routledge.

Chang, C., & Lin, C. J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, *27*, 1–27.

De Martino, F., Valente, G., et al. (2008). Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns. *NeuroImage*, *43*(1), 44–58.

Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *8*, 430–457.

Fombonne, E. (2003). Epidemiological surveys of autism and other pervasive developmental disorders: An update. *Journal of Autism and Developmental Disorders*, *33*(4), 365–382.

Frederick, K. E., Barnard-Brak, L., & Sulak, T. N. (2012). Under-Representation in nationally representative secondary data sets. *International Journal of Research & Method in Education*, *35*(1), 31–40.

Godard, P., McCracken, M., Rollin, J., Van Campen, J., Valdes, K., & Williamson, C. (2007). *Special education elementary longitudinal study: Waves 1, 2, & 3*. Washington, DC: U. S. Office of Special Education Programs (OSEP).

Good, C. D., Johnsrude, I. S., Ashburner, J., Henson, R. N., Friston, K. J., & Frackowiak, R. S. (2001). A voxel-based morphometric study of ageing in 465 normal adult human brains. *NeuroImage*, *14*(1 Pt 1), 21–36.

Gotham, K., Pickles, A., & Lord, C. (2009). Standardizing ADOS scores for a measure of severity in autism spectrum disorders. *Journal of Autism and Developmental Disorders*, *39*, 693–705.

Hall, D., Huerta, M. F., McAuliffe, M. J., & Farber, G. K. (2012). Sharing heterogeneous data: The national database for autism research. *Neuroinformatics*, *10*(4), 331–339.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1–55.

Johnson, S. B., Whitney, G., McAuliffe, M., Wang, H., McCreedy, E., Rozenblit, L., et al. (2010). Med Using global unique identifiers to link autism collections. *Journal of the American Medical Informatics Association*, *17*(6), 689–695.

Lai, M. C., Lombardo, M. V., Pasco, G., Ruigrok, A. N., Wheelwright, S. J., Sadek, S. A., et al. (2011). A behavioral comparison of male and female adults with high functioning autism spectrum conditions. *PLoS One*, *6*(6), e20835.

Lambek, R., Tannock, R., Dalsgaard, S., Trillingsgaard, A., Damm, D., & Thomsen, P. H. (2010). Validating neuropsychological subtypes of ADHD: How do children with and without an executive function deficit differ? *Journal of Child Psychology and Psychiatry*, *51*(8), 895–904.

Lohr, S. (2010). *Sampling: Design and analysis* (2nd ed.). Boston, MA: Brooks/Cole.
Lopez, B. R., Lincoln, A. J., Ozonoff, S., & Lai, Z. (2005). Examining the relationship between executive functions and restricted, repetitive symptoms of autistic disorder. *Journal of Autism and Developmental Disorders*, *35*, 445–460.
Lord, C., Risi, S., Lambrecht, L., Cook, E. H., Jr., Leventhal, B. L., DiLavore, P. C., et al. (2000). The autism diagnostic observation schedule–generic: A standard measure of social and communication deficits associated with the spectrum of autism. *Journal of Autism and Developmental Disorders*, *30*, 205–223.
Marshburn, E. C., & Aman, M. G. (1992). Factor validity and norms for the Aberrant Behavior Checklist in community sample of children with mental retardation. *Journal of Autism and Developmental Disorders*, *22*, 357–373.
Muthén, L. K., & Muthén, B. O. (2008). *MPlus user's guide*. Los Angeles, CA: Muthén & Muthén.
National Database for Autism Research. ndar.nih.gov.
National Longitudinal Transition Study (NLTS-2). (2005). *NLTS-2 data documentation and dictionary introduction*. Washington, D.C: U.S. Office of Special Education Programs (OSEP).
NIH Funds New Program to Investigate Causes and Treatment of Autism. (2007). http://www.nimh.nih.gov/science-news/2007/nih-funds-new-program-to-investigate-causes-and-treatment-of-autism.shtml. Accessed 09 February 2013.
Oliver, C., Petty, J., Ruddick, L., & Bacarese-Hamilton, M. (2011). The associations between repetitive, self-injurious, and aggressive behavior in children with severe intellectual disability. *Journal of Autism and Developmental Disorders*, *37*(1), 1–10.
Piwowar, H. A., Day, R. S., & Fridsma, D. B. (2007). Sharing detailed research data is associated with increased citation rate. *PLos One*, *2*(3), e308.
Poline, J. B., Breeze, J. L., Ghosh, S., Gorgolewski, K., Halchenko, Y. O., Hanke, M., et al. (2012). Data sharing in neuroimaging research. *Frontiers in Neuroinformatics*, *6*, 9.
Richards, C., Oliver, C., Nelson, L., & Moss, J. (2012). Self-injurious behaviour in individuals with autism spectrum disorder and intellectual disability. *Journal of Intellectual Disability Research*, *56*(5), 476–489.
Richman, D., Barnard-Brak, L., Bosch, A., Thompson, S., Grubb, L., & Abby, L. (2013). Predictors of self-injurious behavior exhibited by individuals with autism spectrum disorder. *Journal of Intellectual Disability Research*, *57*(5), 429–439.
Savage, C. J., & Vickers, A. J. (2009). Empirical study of data sharing by authors publishing in PLoS Journals. *PLoS One*, *4*(9), e7078.
Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research*, *8*(2), 23–74.
Sharing Data via the National Database for Autism Research. (2009). http://grants.nih.gov/grants/guide/notice-files/NOT-MH-09-005.html.
Smith, E. (2008). *Using secondary data in educational and social research*. London, UK: Open University Press.
Special Education Elementary Longitudinal Study (SEELS), (2005). *SEELS data documentation and dictionary: Introduction*. Washington, D. C: U.S. Office of Special Education Programs (OSEP).
Stoel, R. D., van den Wittenboer, G., & Hox, J. J. (2003). Methodological issues in the application of the latent growth curve model. In K. van Montfort, H. Oud, & A. Satorra (Eds.), *Recent developments in structural equation modeling: Theory and applications*. Amsterdam: Kluwer Academic Press.
Supekar, K., Uddin, L., & Menon, V. (2013). Brain basis of autism in girls. In *Preliminary results published presented at the 2013 International Meeting for Autism Research, Spain*, Final manuscript: Under review.

Turner, M. (1999). Annotation: Repetitive behaviour in autism: A review of psychological research. *Journal of Child Psychology and Psychiatry*, *40*, 839–849.

Uddin, L. Q., Menon, V., Young, C. B., Ryali, S., Chen, T., Khouzam, A., et al. (2011). Multivariate searchlight classification of structural magnetic resonance imaging in children and adolescents with autism. *Biological Psychiatry*, *70*(9), 833–841.

Uddin, L. Q., Supekar, K. S., Lynch, C., Khouzam, A., Phillips, J., Feinstein, C., et al. (2013). Brain network based classification and prediction of symptom severity in children with autism. *Archives of General Psychiatry*, *70*(8), 869–879.

Weston, R., & Gore, P. A. (2006). A brief guide to structural equation modeling. *The Counseling Psychologist*, *34*(5), 719–751.