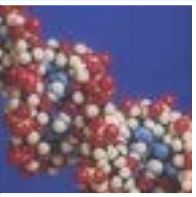
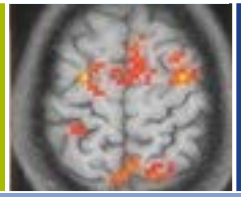
Data Integration and Data Management

The NIMH Data Archive

Greg Farber

Office of Technology Development and Coordination
National Institute of Mental Health





The National Database for Autism Research has Transformed and Expanded into the NIMH Data Archives (NDA)

National Institute of Mental Health

Why Bother Aggregating Data?

1) Understanding complex conditions requires data from large numbers of subjects.
   - Genetic studies have shown that tens of thousands of subjects are required for a partial understanding of the genes associated with neurological diseases.
   - When environmental influences are also important in understanding a disorder, the numbers of subjects needed are likely to be much larger.
   - In addition to requiring large numbers of subjects, understanding complex conditions also requires aggregating many different types of data in a meaningful way.
2) Aggregating data from different laboratories allows the research community to understand how similar (or not) the data being collected really are. This leads to agreement on the best way to perform certain experiments. (common data elements)
3) Depositing data to a repository on a regular basis during data collection allows the laboratory to improve the rigor and reproducibility of their experiments.
4) Aggregating data allows the research community to evaluate the costs and outcomes from different ways of collecting data.

Who Contributes Data to the NDA?

- NIMH awardees who are doing experiments using human subjects
  - National Database for Autism Research
  - Legacy clinical trials funded by NIMH
  - Data from all applications submitted to NIMH after May 1, 2015
  - Pediatric MRI Study
  - A total of 550 awardees are currently expected to share data
- Data from the Adolescent Brain Cognitive Development Study
- Data from the Human Connectome Project (coming soon)
- Awards made by other funding agencies
  - Stanley Foundation
  - Autism Science Foundation
- NDA is federated with
  - Autism Tissue Program
  - Autism Genetic Research Exchange
  - Interactive Autism Network
  - Simons Foundation Autism Research Initiative
  - Ontario Brain Institute (in progress)

NDA Overview

- NDA is a federal data repository that accepts data from the research community.
- The NDA only contains data from human subjects. We have some capability to deal with data that has different types of consent, but NIMH funded data is broadly consented for use by the research community.
- NIMH data are available to the research community through a not too difficult application process that involves a data access committee. (Currently support 4 independent DACs.)
- Summary data are available to everyone with a browser at https://data-archive.nimh.nih.gov/

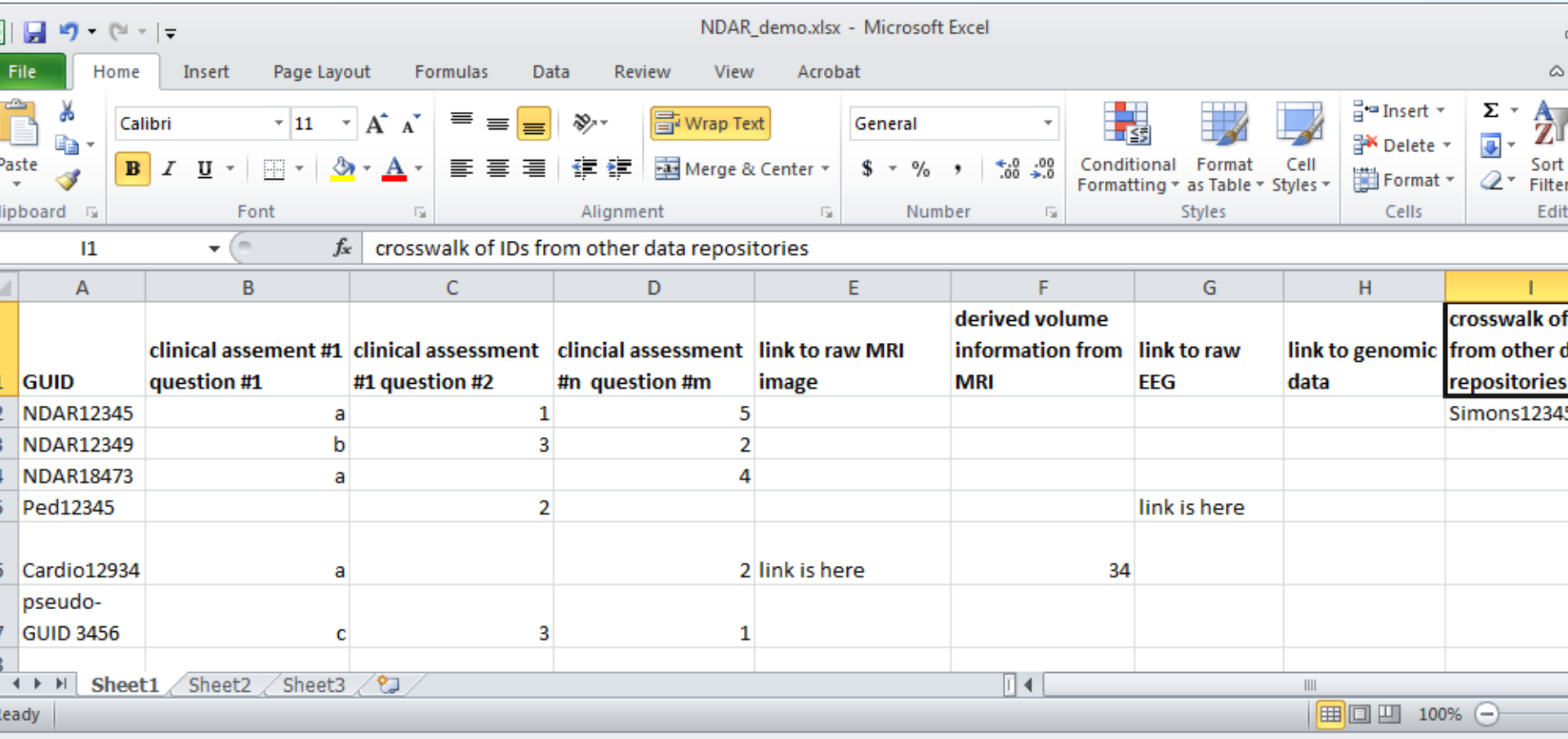
NDA – Types of Data (January 2017)

Type of Data	Participants Submitted	Participants Shared
Any	146,274	131,314
MRI	7,154	4,824
Eye Tracking	1,309	723
Genomics	34,903	32,119
EEG	4,292	803

- ~800 terabytes of imaging, –omic, and other complex experimental data is secured in the Amazon cloud. We expect this go grow to 4-5 petabyte in the next 5 years.
- (kilobyte < megabyte < gigabyte < terabyte < petabyte)
- (document file < PowerPoint presentation < movie < Library of Congress < 13 year long movie)

NDA Basic Structure

- NDA can be thought of as a large two dimensional matrix.
- Dimension 1: The data dictionaries which provide definitions for clinical assessments, imaging experiments, or any other experimental data are the other dimension of the matrix.
- Dimension 2: Global Unique Identifiers (GUID) which are generated using personally identifiable information at the research site are one dimension of the matrix. The GUIDs allow data from the same subject who was seen in different laboratories to be aggregated without requiring that the NDA have any personally identifiable information.
- A variety of queries have been implemented to allow researchers to find the data they are interested in. The queries cross all of the parts of the NDA and also reach into other federated data repositories outside of NIMH.



	A	B	C	D	E	F	G	H	I
1	GUID	clinical assement #1 question #1	clinical assessment #1 question #2	clincial assessment #n question #m	link to raw MRI image	derived volume information from MRI	link to raw EEG	link to genomic data	crosswalk of from other d repositories
2	NDAR12345	a	1	5					Simons12345
3	NDAR12349	b	3	2					
4	NDAR18473	a		4					
5	Ped12345		2				link is here		
6	Cardio12934	a		2	link is here	34			
7	pseudo-GUID 3456	c	3	1					
8									

National Institute of Mental Health

Data Dictionary – The First Building Block

- The NDA data dictionary is one of the key building blocks for this repository. It provides a flexible framework that allows us to work with the research community to define the data they are collecting.
- 1500+ data collection instruments (measures, forms) which are freely available to anyone
  - 130,000+ unique data elements (“questions”) and growing
  - A research community platform for defining the complex language characterizing mental health research
    - Clinical
    - Genomics/Proteomics
    - MRI Modalities
    - Other complex data (EEG, Eye Tracking)
- Accommodates any data type and data structure
- Curated by NDA Staff
- Allows investigators to quickly perform quality control tests of their data without submitting data anywhere by validating that the answer to each data element is within an expected range.

Data Dictionary List (1500+ Measures)

Home | Query | Harmonization Tools | Cloud | Contribute | Request Access | Policy | Tutorials | About | FAQ | Tools | login

Query Data | Data from Labs | Data from Papers | Query by Data Dictionary | Query by Concept | Query by GUID | Query Instructions

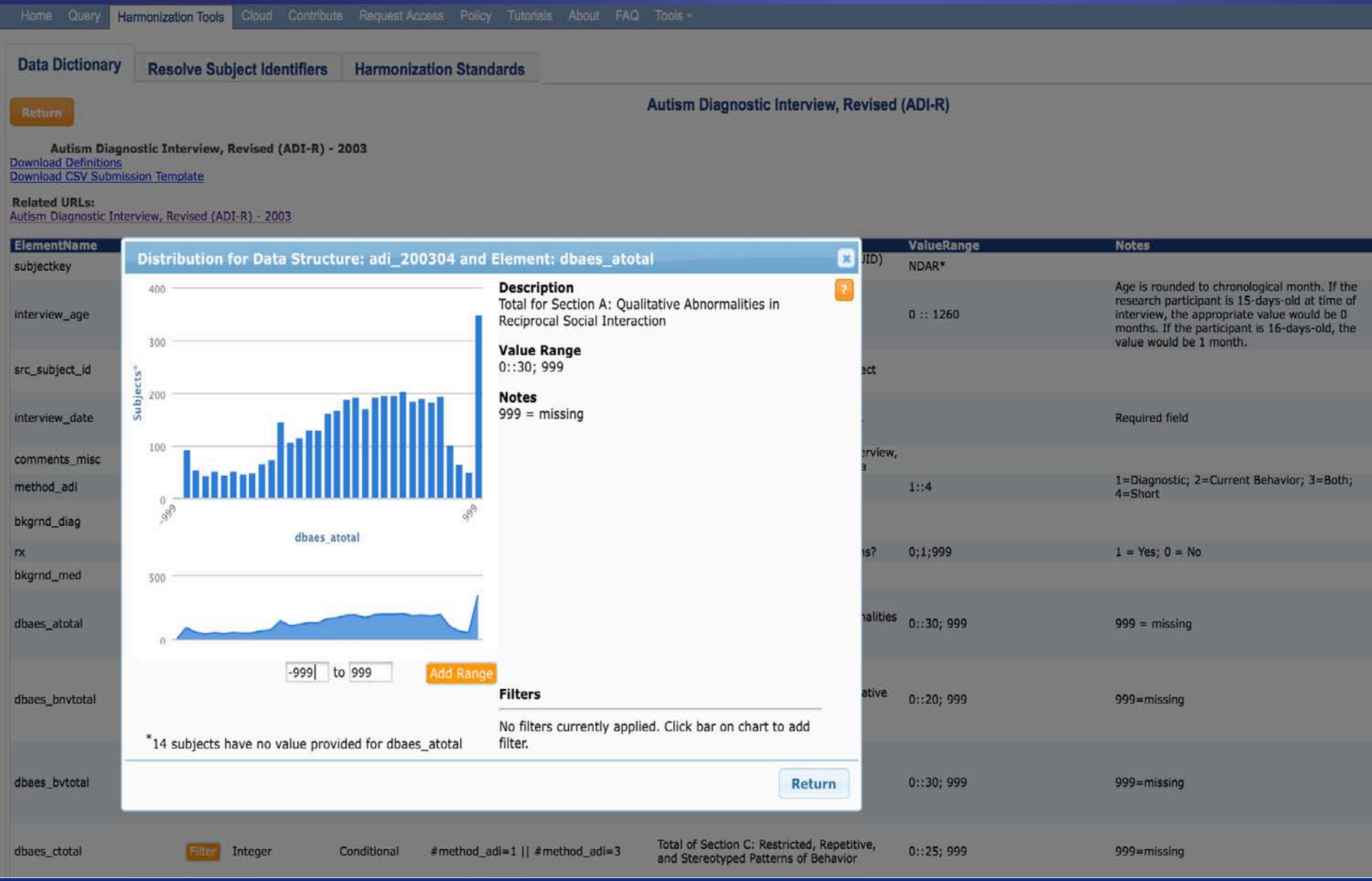
Listed below are the data structures supporting NDAR's autism data definition. To see other definitions in NDAR, select Source. Select Category to see the different types of data structures now available.

Type: All | Source: NDAR | Category: All

ADHD, Acoustics, Adverse Events, Aggression, Anxiety, Behavior, Cognitive, Conflict, Coping, DTI, MRI, fMRI, Demographics, Depression, Diagnostic, EEG, EGG, EMG, ERP, Emotions, Evaluated Data, Exposure, Eye Tracking, Fear, Food, Gen Test, IQ, Life Events, Loneliness, MEG, Med History, Network, OCD, Omics, Personality, Phobia, Phys Characteristics

DOWNLOAD	FILTER	TITLE	SHORT NAME	SOURCE	CATEGORY	SUBMISSION
Download	Filter	A Developmental NEuroPSYchological Assessment	nepsy01	NDAR	Cognitive	Allowed
Download	Filter	ACE Family Medical History	ace_fammedhist01	ACE Common Measures V2, NDAR	Med History	Allowed
Download	Filter	ACE Subject Medical History	ace_subjmedhist01	ACE Common Measures V2, NDAR	Med History	Allowed
Download	Filter	ACE Subject Physical Exam	ace_physexam01	ACE Common Measures V2, NDAR	Phys Exam	Allowed
Download	Filter	ADHD Rating Scale	adhdrs01	NDAR	ADHD	Allowed
Download	Filter	AIR Self-Determination Scale	airsds01	NDAR	Questionnaire	Allowed
Download	Filter	Aberrant Behavior Checklist (ABC) - Community	abc_community02	NDAR, NDCT	Behavior	Allowed
Download	Filter	Abnormal Involuntary Movement Scale	aims01	NDAR, NDCT, RDoC	Questionnaire	Allowed
Download	Filter	Academic Support Scale	asups01	NDAR	Questionnaire	Allowed
Download	Filter	Acceptability Questionnaire	acquest01	NDAR	Questionnaire	Allowed
Download	Filter	Adaptation Phase Protocol and Interview	appi01	NDAR	Questionnaire	Allowed
Download	Filter	Adapted ADOS Module 1	aados_m101	NDAR	Diagnostic	Allowed
Download	Filter	Adapted ADOS Module 2	aados_m201	NDAR	Diagnostic	Allowed
Download	Filter	Adaptive Behavior Assessment System, Second E	abas01	NDAR	Behavior	Allowed
Download	Filter	Adolescent Symptom Inventory	asi01	NDAR	Questionnaire	Allowed
Download	Filter	Adult Adolescent Parenting Inventory *new*	aapi01	NDAR, RDoC	Questionnaire	Allowed
Download	Filter	Adult Behavior Check List	abcl_men_200301	NDAR	Behavior	Allowed
Download	Filter	Adult Impairment Rating Scale	airs01	NDAR, NDCT	Questionnaire	Allowed
Download	Filter	Advanced Normalization Tools (ANTs) Cortical Thi	antsvol01	NDAR	Evaluated Data	Allowed
Download	Filter	Adverse Events	adev01	NDAR, NDCT	Adverse Events	Allowed
		Advocacy Form	advoc01	NDAR	Questionnaire	Allowed
Download	Filter	Age Differentiation Test	adt3601	NDAR	Task Based	Allowed

Data Inspection – Available to All

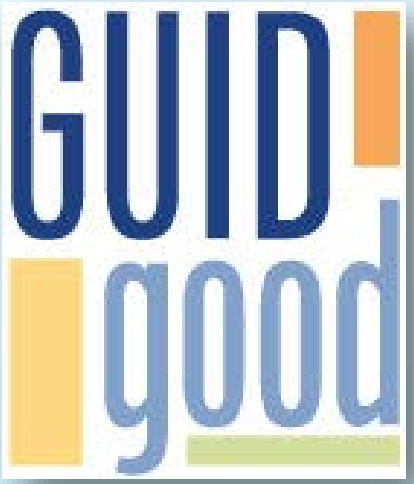


The Data Dictionary is a key component of improving rigor and reproducibility

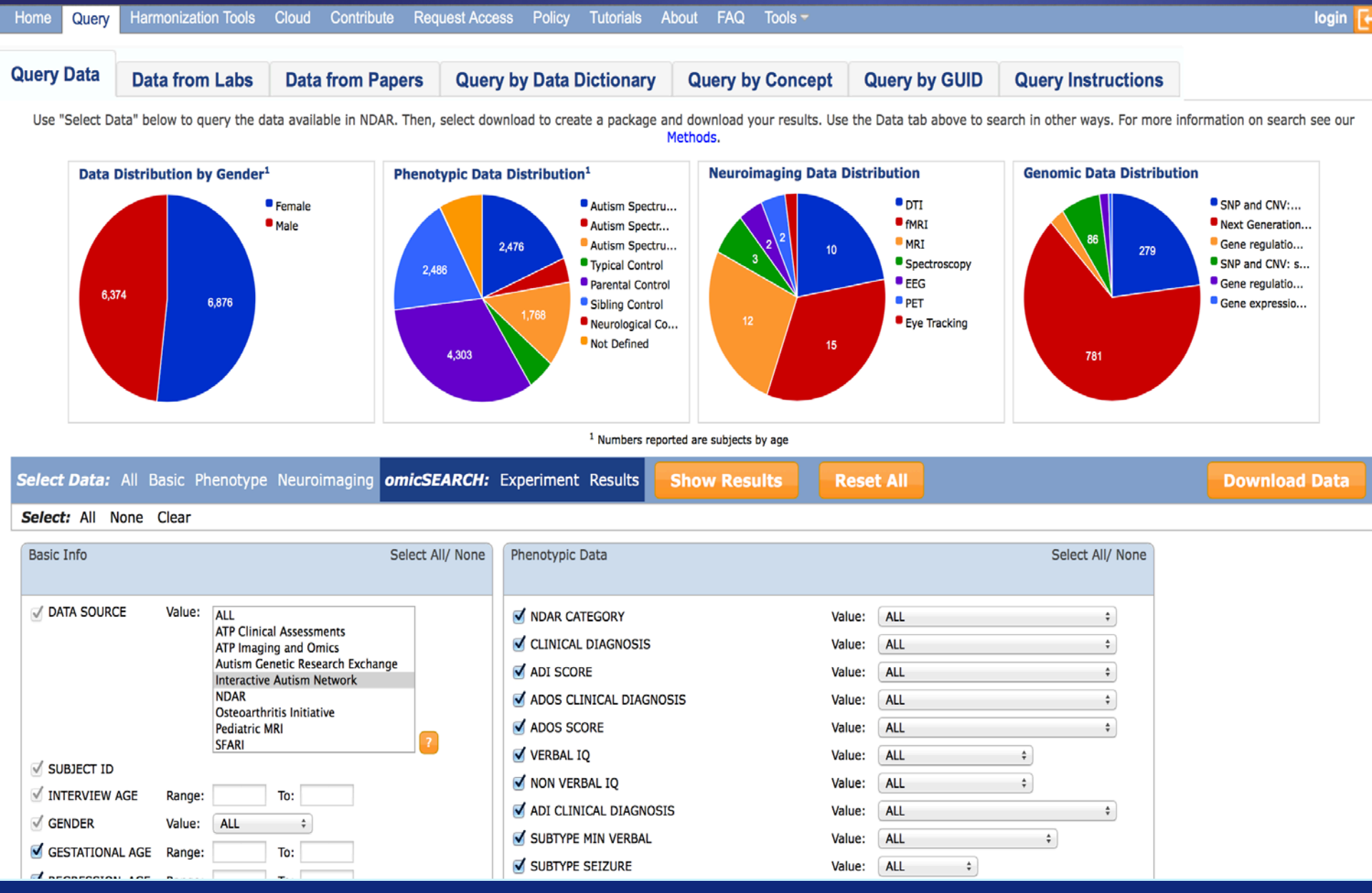
- NDA makes a validation tool available to all, so that if a data dictionary exists, anyone can test their data using the tool to make sure that the recorded information for a subject is consistent with the allowed values in the data dictionary.
- The large number of data dictionaries already available as well as our willingness to create additional data dictionaries as necessary makes this validation very useful.

Global Unique Identifier – the Other Building Block

- The NDA GUID software allows any researcher to generate a unique identifier using some information from a birth certificate.
- If the same information is entered in different laboratories, the same GUID will be generated.
- This strategy allows NDA to aggregate data on the same subject collected in multiple laboratories without holding any of the personally identifiable information about that subject.
- NDA also assigns unique identifiers that do not allow data aggregation (pseudo-GUID) in cases where the GUID could not be generated.
- The GUID is now being used in other research communities (see http://www.youtube.com/watch?v=Tb6euCVoous)



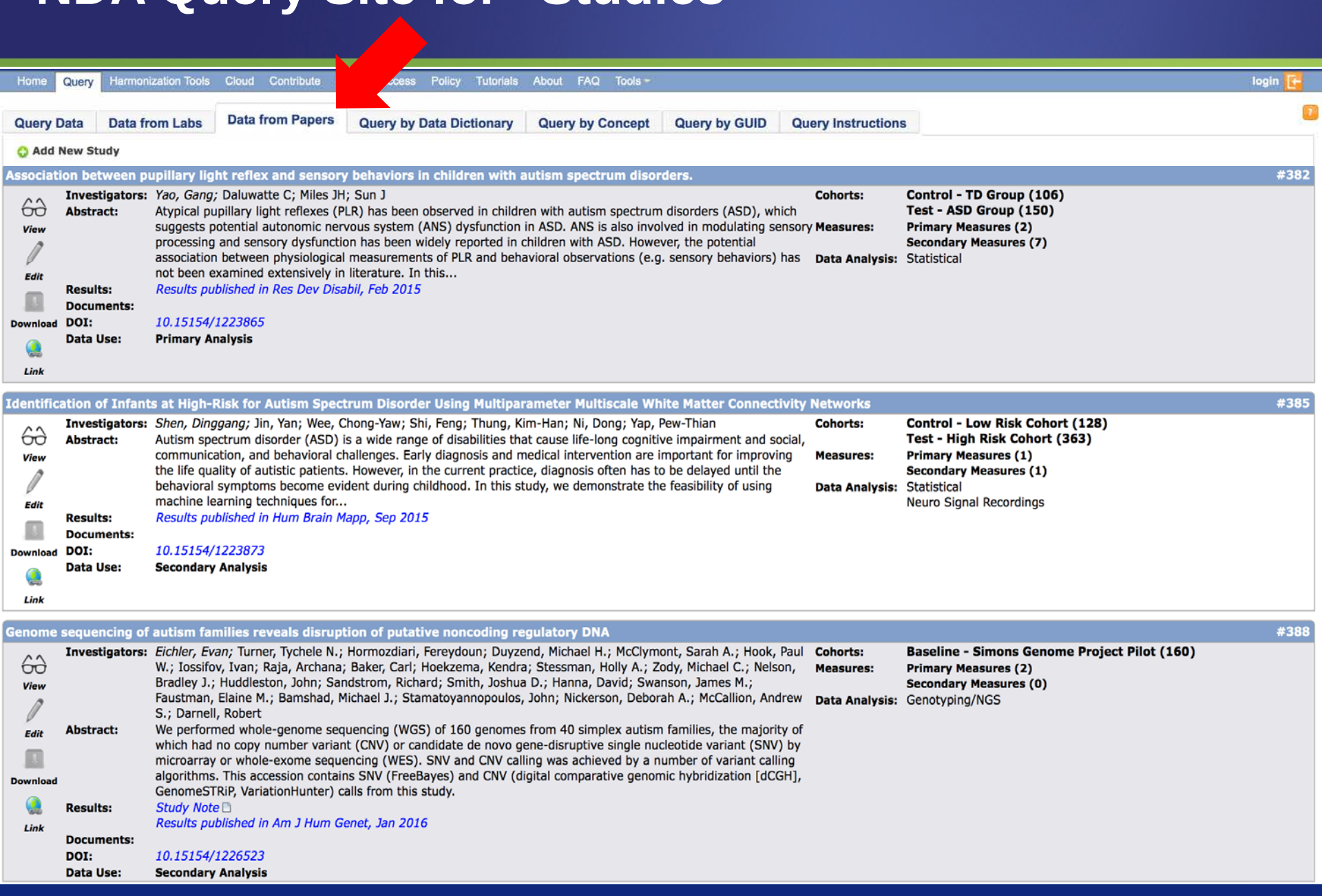
General Query – IAN Example – GUID Works



Data Archives also Allow Data to be Aggregated in Ways not Anticipated by those who Measured the Data

- The NDA allows a user to aggregate data into a “study”.
- The data could all come from a single laboratory, or could come from a variety of sources.
- Digital Object Identifiers are assigned to each study, so it is very easy for an author to deposit data into the NDA and then get a unique identifier that can be referenced in a publication.
- The NDA is happy to accept any data related to mental illness (broadly defined), so the archive does provide a data storage infrastructure that could be useful for many journals.

NDA Query Site for “Studies”



Details of a Study Showing the doi as well as the source of the data from 3 different laboratories

Home Query Harmonization Tools Cloud Contribute Request Access Policy Tutorials About FAQ Tools login

Owner: Niklas Krumm **Owner E-mail:** *nkrumm@uw.edu* **State:** Shared

Transmission disequilibrium of small CNVs in simplex autism. #312

Summary | *Cohorts (3)* | *Measures (6)* | *Types* | *Data Analysis*

Investigators: *Eichler EE;* Krumm N, O'Roak BJ, Karakoc E, Mohajeri K, Nelson B, Vives L, Jacquemont S, Munson J, Bernier R

Abstract: Cohorts: 411 ASD Quads from Simons Simplex Collection 177 Quads from Sanders et al. (PubMed ID: 22495306) 166 Quads from I. Iossifov et al. (PubMed ID: 22542183) 71 Quads from O'Roak et al. (PubMed ID: 22495309) Publication Abstract: We searched for disruptive, genic rare copy-number variants (CNVs) among 411 families affected by sporadic autism spectrum disorder (ASD) from the Simons Simplex Collection by using available exome sequence data and CoNIFER (Copy Number Inference from Exome Reads). Compared to high-density SNP microarrays, our approach yielded ¿2× more smaller genic rare CNVs. We found that affected probands inherited more CNVs than did their siblings (453 versus 394, p = 0.004; odds ratio [OR] = 1.19) and that the probands' CNVs affected more genes (921 versus 726, p = 0.02; OR = 1.30). These smaller CNVs (median size 18 kb) were transmitted preferentially from the mother (136 maternal versus 100 paternal, p = 0.02), although this bias occurred irrespective of affected status. The excess burden of inherited CNVs among probands was driven primarily by sibling pairs with discordant social-behavior phenotypes (p < 0.0002, measured by Social Responsiveness Scale [SRS] score), which contrasts with families where the phenotypes were more closely matched or less extreme (p > 0.5). Finally, we found enrichment of brain-expressed genes unique to probands, especially in the SRS-discordant group (p = 0.0035). In a combined model, our inherited CNVs, de novo CNVs, and de novo single-nucleotide variants all independently contributed to the risk of autism (p < 0.05). Taken together, these results suggest that small transmitted rare CNVs play a role in the etiology of simplex autism. Importantly, the small size of these variants aids in the identification of specific genes as additional risk factors associated with ASD.

Results: *Results published in Am J Hum Genet, Oct 2013*

Documents:

DOI: *10.15154/1163542*

Data Use: Secondary Analysis

Cohorts: **Control - Parental Controls (822)**
Age: 0 to 1,200 months
Gender: Both
Control - Probands (411)
Age: 0 to 1,200 months
Gender: Both
Control - Sibling Controls (411)
Age: 0 to 1,200 months
Gender: Both

Measures: **Primary Measures: (3)**
Secondary Measures: (3)

Data Analysis: Genotyping/NGS
Type: CNV, Exome Sequencing
NGS Analysis
Variant filtering: De novo variants, Inherited variants
Variant validation: Targeted aCGH
Test Statistics: Pearson's chi-square test, Fisher's exact test
Test Correction: FDR
Software: CoNIFER, mrsFAST
Statistical
Method: Student's t-test, Binomial test

Attribution Report

ID	Collection	Subjects
1878	*Genomic Identification of Autism Loci*	1,644
1936	*Deep sequencing of autism candidate genes in 2000 families from the Simons Simplex Collection (SSC)*	462
1895	*Genomic Profiling and Functional Mutation Analysis in Autism Spectrum Disorders*	53

NIMH Data Archives Staff

National Institute of Mental Health

NDA Summary

The NDA is a useful data archive that makes human subjects data:

A) Discoverable – federation, useful queries, XML web services

B) Useful to Others – data access, data QC, data analysis pipelines, APIs

C) Citable – data from labs that conduct experiments, data from papers, dois for groupings of data

D) Linked to the Literature – data link in PubMed as well as data dois in specific publications