NDAR Big Data Realities

Dan Hall

Manager, National Database for Autism Research. NIMH

NIH National Institute of Mental Health

Why Should Anyone Care About Big Data?

- Genomics studies have allowed great strides in uncovering the basis for a number of diseases that are caused by a mutation to a single gene.
- For complex diseases (multiple genes involved and may have an environmental component as well), neither genomics nor imaging (MRI) have been very successful at finding biomarkers that have a strong correlation to diagnosis or treatment.
- For many complex diseases a single diagnosis (diabetes, schizophrenia…) really covers a wide range of diseases that share some symptoms.

Potential Solutions:

- Continued experiments to uncover the basic biology.
- Combining data from experiments in multiple laboratories with continuous aggregation of data across modalities and results

National Database For Autism Research

- NDAR is trying to solve a hard problem:
  - The data in NDAR come from multiple laboratories and are measured for different purposes.
  - There are many data collection instruments that measure similar clinical criteria.
  - There are a number of cases where the same individual is seen in multiple different laboratories.
  - There are other significant databases with autism data.

National Database for Autism Research

- Joint initiative supported by NIMH, NICHD, NINDS, and NIEHS
  - Federal data repository
  - Contains data from human subjects related to autism (and control subjects)
  - Data are available to the research community through a not too difficult application process
  - Summary data are available to everyone with a browser at http://ndar.nih.gov
- Begun in late 2006, first data was received in 2008, significant data became available in 2012.
- Autism Interagency Coordinating Committee Recommendation called for 90% of all human subjects data to be shared.
- Currently has broad data available from over 77,000 subjects from demographic data, -omic (~400TB), clinical assessments, imaging data, eye tracking, exposure

NDAR: Implementation

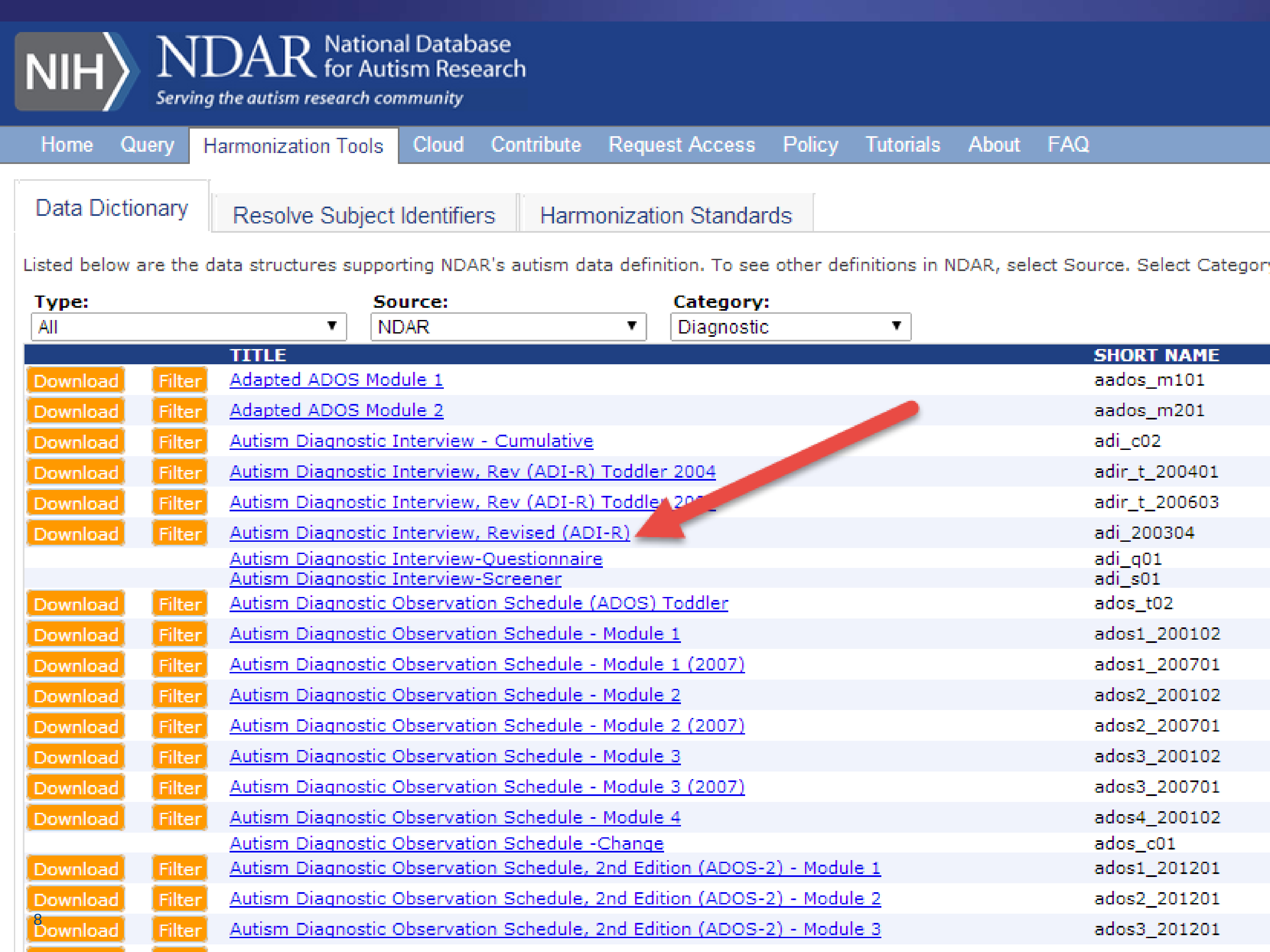
- NDAR has deep federation with the following data repositories allowing NDAR to query data controlled by others.
  - Autism Tissue Program
  - Autism Genetic Resource Exchange
  - Interactive Autism Network
  - Simons Foundation Autism Research Initiative
- Generally, NIH funded investigators are expected to share their data via NDAR ongoing for data **"about"** research subjects and at time of publication on **"findings"**. Investigators with funding from other sources are welcome to deposit their data.
- Over 100 studies have registered data, and more than 150 are expected to share data.
- NDAR has two key features to allow data standardization and aggregation: data dictionaries and the Global Unique Identifier (GUID)

Global Unique Identifier

- The NDAR 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 NDAR to aggregate data on the same subject collected in multiple laboratories without holding any of the personally identifiable information about that subject.
- The GUID is now being used in other research communities and can be made available to you.

Data Dictionary

- The NDAR data dictionary is one of the key building blocks for this repository. It provides a flexible and extensible framework for data definition by the research community.
- 400+ instruments, freely available to anyone
  - 60,000+ unique data elements and growing
  - A research community platform for defining the complex language characterizing autism research
    - Clinical
    - Genomics/Proteomics
    - Imaging Modalities
- Accommodates any data type and data structure
- Extended and enhanced by the ASD research community
- Curated by NDAR
- Allows investigators to quickly perform quality control tests of their data without submitting data anywhere.



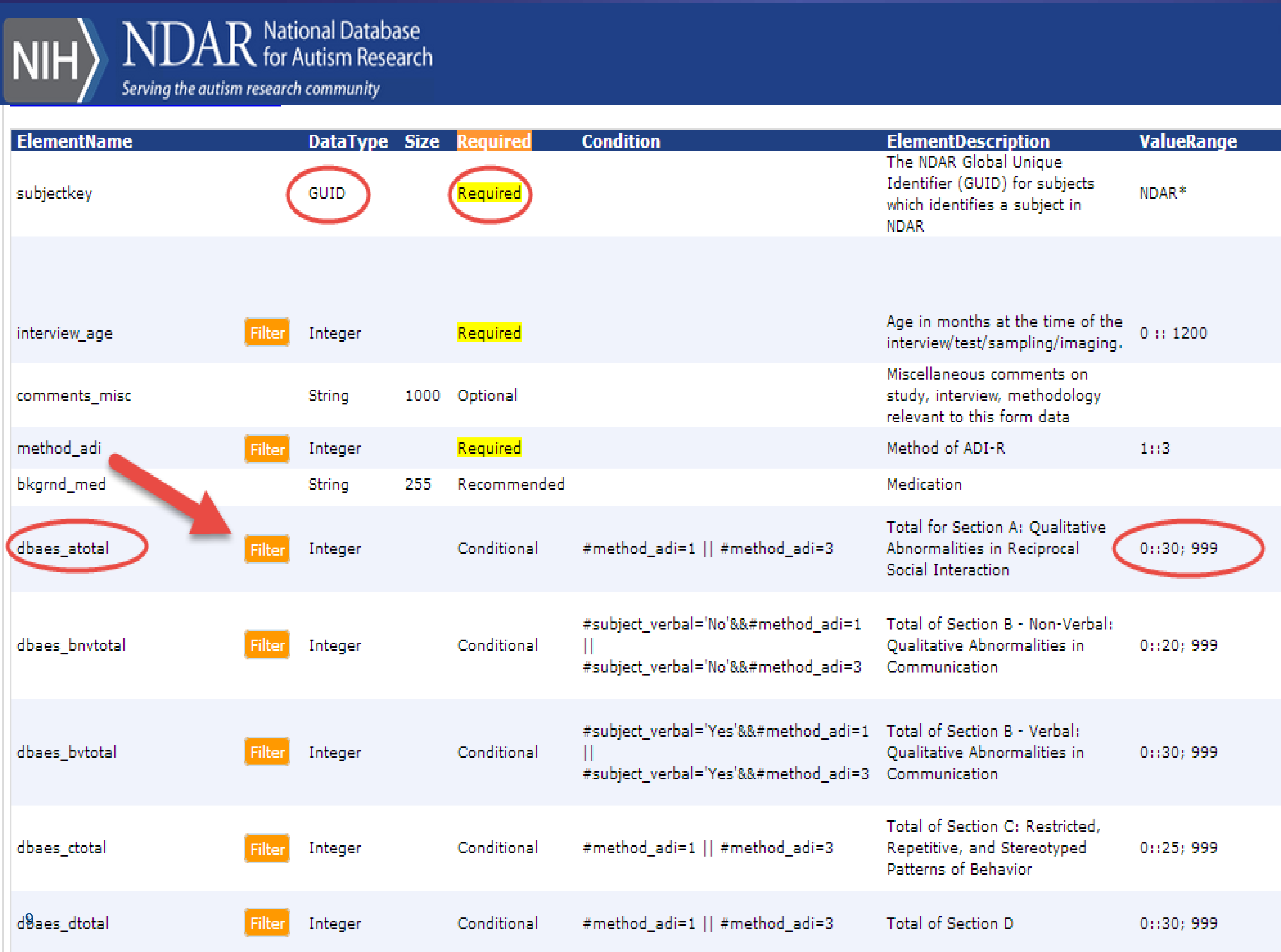
Listed below are the data structures supporting NDAR's autism data definition. To see other definitions in NDAR, select Source. Select Categor

Type: All

Source: NDAR

Category: Diagnostic

		TITLE	SHORT NAME
Download	Filter	Adapted ADOS Module 1	aados_m101
Download	Filter	Adapted ADOS Module 2	aados_m201
Download	Filter	Autism Diagnostic Interview - Cumulative	adi_c02
Download	Filter	Autism Diagnostic Interview, Rev (ADI-R) Toddler 2004	adir_t_200401
Download	Filter	Autism Diagnostic Interview, Rev (ADI-R) Toddler 2006	adir_t_200603
Download	Filter	Autism Diagnostic Interview, Revised (ADI-R)	adi_200304
		Autism Diagnostic Interview-Questionnaire	adi_q01
		Autism Diagnostic Interview-Screener	adi_s01
Download	Filter	Autism Diagnostic Observation Schedule (ADOS) Toddler	ados_t02
Download	Filter	Autism Diagnostic Observation Schedule - Module 1	ados1_200102
Download	Filter	Autism Diagnostic Observation Schedule - Module 1 (2007)	ados1_200701
Download	Filter	Autism Diagnostic Observation Schedule - Module 2	ados2_200102
Download	Filter	Autism Diagnostic Observation Schedule - Module 2 (2007)	ados2_200701
Download	Filter	Autism Diagnostic Observation Schedule - Module 3	ados3_200102
Download	Filter	Autism Diagnostic Observation Schedule - Module 3 (2007)	ados3_200701
Download	Filter	Autism Diagnostic Observation Schedule - Module 4	ados4_200102
		Autism Diagnostic Observation Schedule -Change	ados_c01
Download	Filter	Autism Diagnostic Observation Schedule, 2nd Edition (ADOS-2) - Module 1	ados1_201201
Download	Filter	Autism Diagnostic Observation Schedule, 2nd Edition (ADOS-2) - Module 2	ados2_201201
Download	Filter	Autism Diagnostic Observation Schedule, 2nd Edition (ADOS-2) - Module 3	ados3_201201



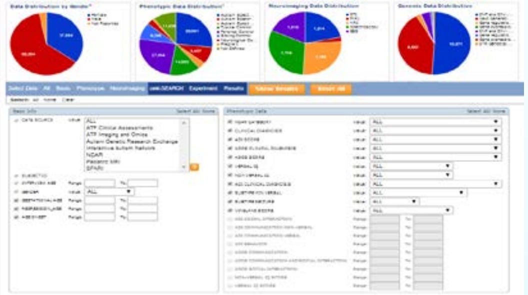
NIH NDAR National Database for Autism Research

Serving the autism research community

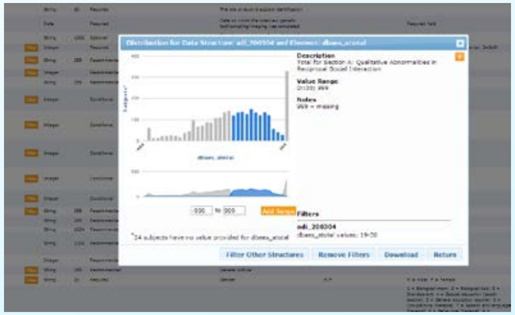
ElementName		DataType	Size	Required	Condition	ElementDescription	ValueRange
subjectkey		GUID		Required		The NDAR Global Unique Identifier (GUID) for subjects which identifies a subject in NDAR	NDAR*
interview_age	Filter	Integer		Required		Age in months at the time of the interview/test/sampling/imaging.	0 :: 1200
comments_misc		String	1000	Optional		Miscellaneous comments on study, interview, methodology relevant to this form data	
method_adi	Filter	Integer		Required		Method of ADI-R	1::3
bkgrnd_med		String	255	Recommended		Medication	
dbaes_atotal	Filter	Integer		Conditional	#method_adi=1 || #method_adi=3	Total for Section A: Qualitative Abnormalities in Reciprocal Social Interaction	0::30; 999
dbaes_bnvtotal	Filter	Integer		Conditional	#subject_verbal='No'&&#method_adi=1 || #subject_verbal='No'&&#method_adi=3	Total of Section B - Non-Verbal: Qualitative Abnormalities in Communication	0::20; 999
dbaes_bvtotal	Filter	Integer		Conditional	#subject_verbal='Yes'&&#method_adi=1 || #subject_verbal='Yes'&&#method_adi=3	Total of Section B - Verbal: Qualitative Abnormalities in Communication	0::30; 999
dbaes_ctotal	Filter	Integer		Conditional	#method_adi=1 || #method_adi=3	Total of Section C: Restricted, Repetitive, and Stereotyped Patterns of Behavior	0::25; 999
dbaes_dtotal	Filter	Integer		Conditional	#method_adi=1 || #method_adi=3	Total of Section D	0::30; 999

Methods for Query, Analysis, and Results Reporting

General Query


Data Distribution by Gender
Phenotypic Data Distribution
Neuroimaging Data Distribution
Genomic Data Distribution
Basic Info
Phenotypic Data
DATA SOURCE
ALL
ATP Clinical Assessments
ATP Imaging and Omics
Autism Genetic Research Exchange
Interactive Autism Network
NDAR
Research Kit
SFARI

By Data Element


Distribution for Data Structure: adi_200304 and Element: dbaes_atotal

Description
Total for Section A: Qualitative Abnormalities in Reciprocal Social Interaction

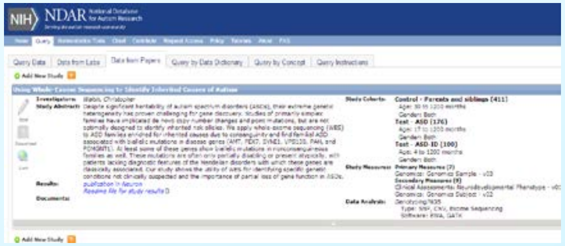
Value Range
0::30; 999

Notes
999 = missing

*24 subjects have no value provided for dbaes_atotal

Filters
adi_200304
dbaes_atotal values: 19-30

From Papers



Using Whole-Exome Sequencing to Identify Inherited Causes of Autism

Investigators: Walsh, Christopher

Study Abstract: Despite significant heritability of autism spectrum disorders (ASDs), their extreme genetic heterogeneity has proven challenging for gene discovery. Studies of primarily simplex families have implicated de novo copy number changes and point mutations, but are not optimally designed to identify inherited risk alleles. We apply whole-exome sequencing (WES) to ASD families enriched for inherited causes due to consanguinity and find familial ASD associated with biallelic mutations in disease genes (AMT, PEX7, SYNE1, VPS13B, PAH and POMGNT1). At least some of these genes show biallelic mutations in nonconsanguineous families as well. These mutations are often only partially disabling or present atypically, with patients lacking diagnostic features of the Mendelian disorders with which these genes are classically associated. Our study shows the utility of WES for identifying specific genetic conditions not clinically suspected and the importance of partial loss of gene function in ASDs.

Results: publication in Neuron
Readme file for study results

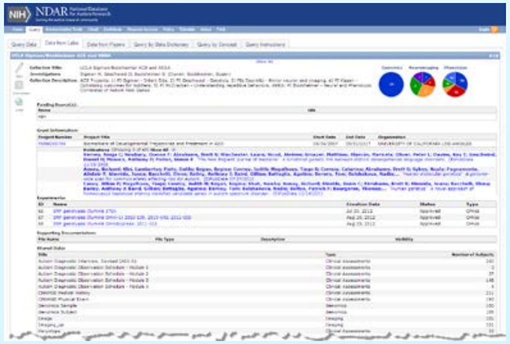
Documents:

Study Cohorts: **Control - Parents and siblings (411)**
Age: 30 to 1200 months
Gender: Both
Test - ASD (176)
Age: 17 to 1200 months
Gender: Both
Test - ASD-ID (100)
Age: 4 to 1200 months
Gender: Both

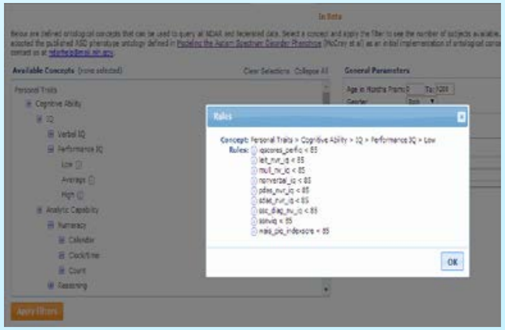
Study Measures: **Primary Measures (7)**
Genomics: Genomics Sample - v03
Secondary Measures (9)
Clinical Assessments: Neurodevelopmental Phenotype - v0
Genomics: Genomics Subject - v02

Data Analysis: Genotyping/NGS
Type: SNP, CNV, Exome Sequencing
Software: BWA, GATK

From Labs



By Concept


Below are defined ontological concepts that can be used to query all NDAR and federated data. Select a concept and apply the filter to see the number of subjects available. ... adopted the published ASD phenotype ontology defined in Modeling the Autism Spectrum Disorder Phenotype (McCray et al) as an initial implementation of ontological con... contact us at ndarhelp@mail.nih.gov.

Available Concepts (none selected)

- Personal Traits
  - Cognitive Ability
    - IQ
      - Verbal IQ
      - Performance IQ
        - Low
        - Average
        - High
    - Analytic Capability
      - Numeracy
        - Calendar
        - Clock/Time
        - Count
      - Reasoning

General Parameters
Age in Months From: 0 To: 1200
Gender: Both

Rules

Concept: Personal Traits > Cognitive Ability > IQ > Performance IQ > Low

Rules:
- iqscores_perfiq < 85
- leit_nvr_iq < 85
- mull_nv_iq < 85
- nonverbal_iq < 85
- pdas_nvr_iq < 85
- sdas_nvr_iq < 85
- ssc_diag_nv_iq < 85
- sbnviq < 85
- wais_pio_indexscore < 85

Does It Work?

- More than 270 investigators have requested access to the NDAR database.
- Papers are starting to appear based largely or solely on data from NDAR. Investigators are also using data from NDAR as preliminary evidence in NIH grant applications.
- The NDAR infrastructure has been cloned to create the Federal Interagency Traumatic Brain Injury informatics system (FITBIR).
- NIMH has recently decided to collect data from clinical trials and from Research Domain criteria expanding the systems scope.

Conclusions

- It is possible, but not easy, to make information from human subjects research broadly available to the research community.
- The GUID and the data dictionary are the key elements that are needed to allow complex queries across data from multiple laboratories.
- Behind the scenes data curation is also needed. This is costly, but the user does not see this effort (unless it fails).
- This Model of Data Sharing and Results Reporting is appropriate to other Research Communities, especially those related to Research Domain Criteria (RDoC).