

# 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)





# Why Bother Aggregating Data?

- ) 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 <u>https://data-archive.nimh.nih.gov/</u>



# 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)</li>
- (document file < PowerPoint presentation < movie < Library of Congress < 13 year long movie)</li>



# **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.



	🚽 🍠 • (°' •	-   -		NDAR.	_demo.xlsx - Microsoft	Excel			
F	ile Home	Insert Page Layo	out Formulas Da	ata Review View	Acrobat				۵
[ ]	Cal			≫ · → Wrap Ter	Kt General	• €.0 .00 Cond	itional Format	Cell	× Σ × A Z Sort
p	board 🖬	Font		Alignment	S Numl	ber G	tting + as Table + S Styles	ityles → Format Cells	<ul> <li>Filter</li> <li>Edit</li> </ul>
	11	$\bullet$ (a) $f_x$	crosswalk of IDs fr	om other data reposi	tories				
ſ	А	В	С	D	E	F	G	Н	1
		clinical assement #1 clinical asse		clincial assessment	link to raw MRI	derived volume information from	link to raw	link to genomic	crosswalk of from other d
GUID		question #1	#1 question #2	#n question #m	image	MRI	EEG	data	repositories
	NDAR12345	a 1 5						Simons1234	
	NDAR12349	b	3	2					
	NDAR18473	а		4					
	Ped12345		2				link is here		
	Cardio12934	а		2	link is here	34	ı		
	pseudo-								
	GUID 3456	с	3	1					
		1 Chaot2 Chaot2							
	adv								0% 🕞



# 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 →
 Iogin
 Iogin

 Query
 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:		Source:	Category:				
All		NDAR	∢ All				
DOWNLOAD	) FILTER	R TITLE	ADHD	SHORT NAME	SOURCE	CATEGORY	SUBMISSION
Download	Filter	A Developmental NEuroPSYchological Assessment	Adverse Events	nepsy01	NDAR	Cognitive	Allowed
Download	Filter	ACE Family Medical History	Aggression	ace_fammedhist01	ACE Common Measures V2,	NDAR Med History	Allowed
Download	Filter	ACE Subject Medical History	Behavior	ace_subjmedhist01	ACE Common Measures V2,	NDAR Med History	Allowed
Download	Filter	ACE Subject Physical Exam	Cognitive Conflict	ace_physexam01	ACE Common Measures V2,	NDAR Phys Exam	Allowed
Download	Filter	ADHD Rating Scale	Coping	adhdrs01	NDAR	ADHD	Allowed
Download	Filter	AIR Self-Determination Scale	DTI, MRI, fMRI Demographics	airsds01	NDAR	Questionnaire	Allowed
Download	Filter	Aberrant Behavior Checklist (ABC) - Community	Depression	abc_community02	NDAR, NDCT	Behavior	Allowed
Download	Filter	Abnormal Involuntary Movement Scale	Diagnostic EEG	aims01	NDAR, NDCT, RDoC	Questionnaire	Allowed
Download	Filter	Academic Support Scale	EGG	asups01	NDAR	Questionnaire	Allowed
Download	Filter	Acceptability Questionnaire	ERP	acquest01	NDAR	Questionnaire	Allowed
Download	Filter	Adaptation Phase Protocol and Interview	Emotions Evaluated Data	appi01	NDAR	Questionnaire	Allowed
Download	Filter	Adapted ADOS Module 1	Exposure	aados_m101	NDAR	Diagnostic	Allowed
Download	Filter	Adapted ADOS Module 2	Eye Tracking Fear	aados_m201	NDAR	Diagnostic	Allowed
Download	Filter	Adaptive Behavior Assessment System, Second E	Food	abas01	NDAR	Behavior	Allowed
Download	Filter	Adolescent Symptom Inventory	Gen Test IQ	asi01	NDAR	Questionnaire	Allowed
Download	Filter	Adult Adolescent Parenting Inventory new	Life Events	aapi01	NDAR, RDoC	Questionnaire	Allowed
Download	Filter	Adult Behavior Check List	MEG	abcl_men_200301	NDAR	Behavior	Allowed
Download	Filter	Adult Impairment Rating Scale	Med History	airs01	NDAR, NDCT	Questionnaire	Allowed
Download	Filter	Advanced Normalization Tools (ANTs) Cortical Thi	OCD	antsvol01	NDAR	Evaluated Data	Allowed
Download	Filter	Adverse Events	Omics Personality	adev01	NDAR, NDCT	Adverse Events	Allowed
		Advocacy Form	Phobia	advoc01	NDAR	Questionnaire	Allowed
Download	Filter	Age Differentiation Test	Phys Characteristics	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 <u>http://www.youtube.com/watch?v=Tb6euCVoous</u>)



# **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"

Home	Query Harmoni	nization Tools Cloud Contribute Ress Policy Tutorials About FAQ Tools -	ogin 💽					
Query D	ata Data fr	rom Labs Data from Papers Query by Data Dictionany Query by Concept Query by CUID Query Instructions						
Query D	ata Data In	query by bata bictionary query by concept query by Gold query instructions						
🗘 Add N	lew Study							
Associati	on between pu	upillary light reflex and sensory behaviors in children with autism spectrum disorders.	#382					
	Investigators:	Vao Cohorte: Control - TD Group (105)						
66	Abstract:	Atvoical pupillary light reflexes (PIR) has been observed in children with autism spectrum disorders (ASD), which Test - ASD Group (150)						
View	Abbitati	suggests potential autonomic nervous system (ANS) dysfunction in ASD. ANS is also involved in modulating sensory Measures: Primary Measures (2)						
		processing and sensory dysfunction has been widely reported in children with ASD. However, the potential Secondary Measures (7)						
		association between physiological measurements of PLR and behavioral observations (e.g. sensory behaviors) has Data Analysis: Statistical						
Edit		not been examined extensively in literature. In this						
	Results:	Results published in Res Dev Disabil, Feb 2015						
	Documents:							
Download	DOI:	10.15154/1223865						
	Data Use:	Primary Analysis						
620								
Link								
Identifica	tion of Infant	ts at High-Risk for Autism Spectrum Disorder Using Multiparameter Multiscale White Matter Connectivity Networks	#385					
Lacitation	Investigators	Shan Diagrandi Jin Yani Wee Chong Yani Shi Sandi Thing Kim Hani Ni Dongi Yan Dew Thian Cohorter Control - Low Bick Cohort (128)						
<u>6</u>	Abetract	Antism spertrum disorder (ASD) is a wide range of disabilities that raise life-long condities impairment and social <b>Conditions Conditions</b> (120)						
View	Abstracti	communication, and behavioral challenges Farly diagnosis and medical intervention are important for improving Measures: Primary Measures (1)						
A		the life quality of autistic patients. However, in the current practice, diagnosis often has to be delayed until the Secondary Measures (1)						
0		behavioral symptoms become evident during childhood. In this study, we demonstrate the feasibility of using Data Analysis: Statistical						
Edit		machine learning techniques for Neuro Signal Recordings						
-	Results:	Results published in Hum Brain Mapp, Sep 2015						
1.00	Documents:							
Download	DOI:	10.15154/1223873						
	Data Use:	Secondary Analysis						
60								
Link								
Genome	sequencing of	f autism families reveals disruption of putative noncoding regulatory DNA	#388					
Genome	Tryestigatory	Sicher Firagi Tirrer Tirchel N.; Hermartini Foroudaus Duvrand Michael H.; McChumant Sarah A.; Hook, David, Cahastar Bacalina - Simons Ganoma Braiast Bilat (150)	#300					
62	investigators:	W: Jossifik, Vian Bai, Archana: Baker, Carl: Hoekzema, Kendra: Stessman, Holly A.: Zody, Michael C.: Nelson, Massurer, Demary Massurer (2)						
View		Bradley J.: Huddiston, John: Sandstrom, Richard: Smith, Joshua D.: Hanna, David: Swanson, James M.: Secondary Measures (0)						
		Faustman, Elaine M.; Bamshad, Michael J.; Stamatoyannopoulos, John; Nickerson, Deborah A.; McCallion, Andrew Data Analysis: Genotyping/NGS						
		S.; Darnell, Robert						
Edit	Abstract:	We performed whole-genome sequencing (WGS) of 160 genomes from 40 simplex autism families, the majority of						
		which had no copy number variant (CNV) or candidate de novo gene-disruptive single nucleotide variant (SNV) by						
1.45		microarray or whole-exome sequencing (WES). SNV and CNV calling was achieved by a number of variant calling						
Download		algorithms. This accession contains SNV (FreeBayes) and CNV (digital comparative genomic hybridization [dCGH],						
	Boculto	Struck when by the by t						
000	Results	Results oubliched in Am 1 Hum Genet. Jan 2016						
Link	Documents:							
	DOI:	10.15154/1226523						
	Data Use:	Secondary Analysis						

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

Home Query	Harmonizat	on Tools C	loud Contribute	Request Ac	cess Policy Tutori	ials About F	FAQ Tools <del>-</del>				login 💽
wner: Niklas Kr	rumm Ow	ner E-mail:	nkrumm@uw.edu	State: S	Shared					Transmission disequilibrium of small CNVs in simp	ex autism. #312
Summary	Cohorts	(3) Me	easures (6)	Types	Data Analysis						
Investigators:Eichler EE; Krumm N, O'Roak BJ, Karakoc E, Mohajeri K, Nelson B, Vives L, Jacquemont S, M Bernier RAbstract:Cohorts: 411 ASD Quads from Simons Simplex Collection 177 Quads from Sanders et al. (Pu 22495306) 166 Quads from I. Iossifov et al. (PubMed ID: 22542183) 71 Quads from O'Roak (PubMed ID: 22495309) Publication Abstract: We searched for disruptive, genic rare copy-nu variants (CNVs) among 411 families affected by sporadic autism spectrum disorder (ASD) fro Simons Simplex Collection by using available exome sequence data and CoNIFER (Copy Num Inference from Exome Reads). Compared to high-density SNP microarrays, our approach yiel more smaller genic rare CNVs. We found that affected probands inherited more CNVs than di siblings (453 versus 394, p = 0.004; odds ratio [OR] = 1.19) and that the probands' CNVs al more genes (921 versus 726, p = 0.02; OR = 1.30). These smaller CNVs (median size 18 kb transmitted preferentially from the mother (136 maternal versus 100 paternal, p = 0.02), all this bias occurred irrespective of affected status. The excess burden of inherited CNVs amony 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 familie the phenotypes were more closely matched or less extreme (p > 0.5). Finally, we found enrit brain-expressed genes unique to probands, especially in the SRS-discordant group (p = 0.00 combined model, our inherited CNVs, de novo CNVs, and de novo single-nucleotide variants independently contributed to the risk of autism (p < 0.05). Taken together, these results sug small transmitted rare CNVs play a role in the etiology of simplex autism. Importantly, the sr of these variants aids in the identification of specific genes as additional risk factors associat ASD					Munson J, PubMed ID: ak et al. number from the imber ielded ¿2× did their affected kb) were although ong (p < lies where irichment of 0035). In a is all uggest that small size ated with	Cohorts: Measures: Data Analysis	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 Primary Measures: (3) Secondary Measures: (3) Secondary Measures: (3) Secondary Measures: (3) Secondary Measures: (3) Secondary Measures: (3) Secondary Measures: (3) Variant filtering: De novo variants, Inherited variants 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				
Results: Documents: DOI: (i) Data Use: (i)	<i>Re</i> . 10. Sec	Results published in Am J Hum Genet, Oct 2013 10.15154/1163542 Secondary Analysis									
Attribution Re	eport 🗹	ID			Collection			Subjects			
	1	78 Genom	ic Identification of A	Autism Loci				1,644			
	19	36 Deep si Collecti	equencing of autism ion (SSC)	n candidate g	genes in 2000 families	s from the Simo	ons Simplex	462			
	18	95 Genom	ic Profiling and Fund	ctional Mutat	tion Analysis in Autisn	n Spectrum Disc	orders	53			

# **NIMH Data Archives Staff**





# **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



