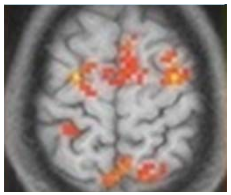
The NIMH Data Sharing Policy

Greg Farber

Director, Office of Technology Development and Coordination
National Institute of Mental Health
farberg@mail.nih.gov

September 2019



Introduction – Many Expectations/Guide Notices Related to Data Sharing

- NIMH has been funding a data archive for human subjects data (NIMH Data Archive, NDA, https://nda.nih.gov/) since 2006.

- Notices related to NIMH data sharing expectations are about as old:
  - NOT-MH-06-108 – data sharing related to bio-samples
  - NOT-MH-09-005 – data sharing related to autism research
  - NOT-MH-14-015 – data sharing related to clinical trials
  - NOT-MH-15-012 – data sharing related to clinical research

- In addition, NIMH researchers have always been expected to follow the NIH wide Genomic Data Sharing Policy.

Introduction – GDS and Streamlining Existing Notices

- The change in the availability of SRA for large scale genomic data has resulted in NIMH deciding to use the NIMH Data Archive to hold all genomic data from human subjects.
- Other ICs have made similar changes with where they are storing genomic data (NOT-HG-19-024, NOT-AA-19-020).

- NOT-MH-19-033 replaces all of the existing data sharing notices and explains to the research community how to comply with GDS (human subjects) for NIMH funded research.
- With the exception of genomic data, NOT-MH-19-033 does not substantially change what NIMH was previously doing.

- Awardees responding to BRAIN Initiative FOAs are subject to a very similar Notice related to BRAIN Initiative data sharing (NOT-MH-19-010).

NOT-MH-19-033 Overview

- Applies to all grant applications that involve human subjects research submitted after January 1, 2020 with a few exceptions (F, K, T, SBIR, R03, R25, AIDS applications).
- Expects all applications to include a resource sharing plan that covers:
  - A summary of the data to be shared
  - The data standards/data dictionaries that will be used to collect the data
  - The schedule to validate the data against the standards/data dictionaries (NDA provides a useful validation tool that can be run without submitting data)
- Awardees will submit data (not share it with the research community) every 6 months to allow the awardee to perform validation checks close to the time when data were measured.
- Reminds the research community that they need to fill out section C.5 “Other Products and Resource Sharing” in the RPPR.

Motivation - 1

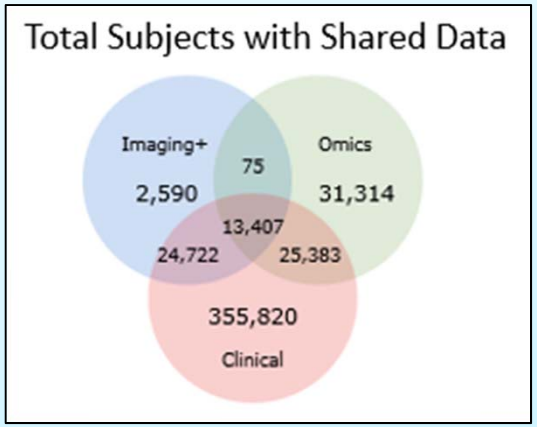
- Why does NIMH care about data sharing?
- The conditions that NIMH is working on are almost all “complex disorders” where the relationship between the underlying biology and the symptoms is unclear.
- There are many people with a particular diagnosis who would be assigned to different subgroups (in that initial diagnostic category or perhaps in a different one) if we understood the biology.
- Aggregating data from large populations may be a useful (necessary?) way to figure out how many subgroups there are inside our heterogeneous diagnostic groups.

Motivation - 2

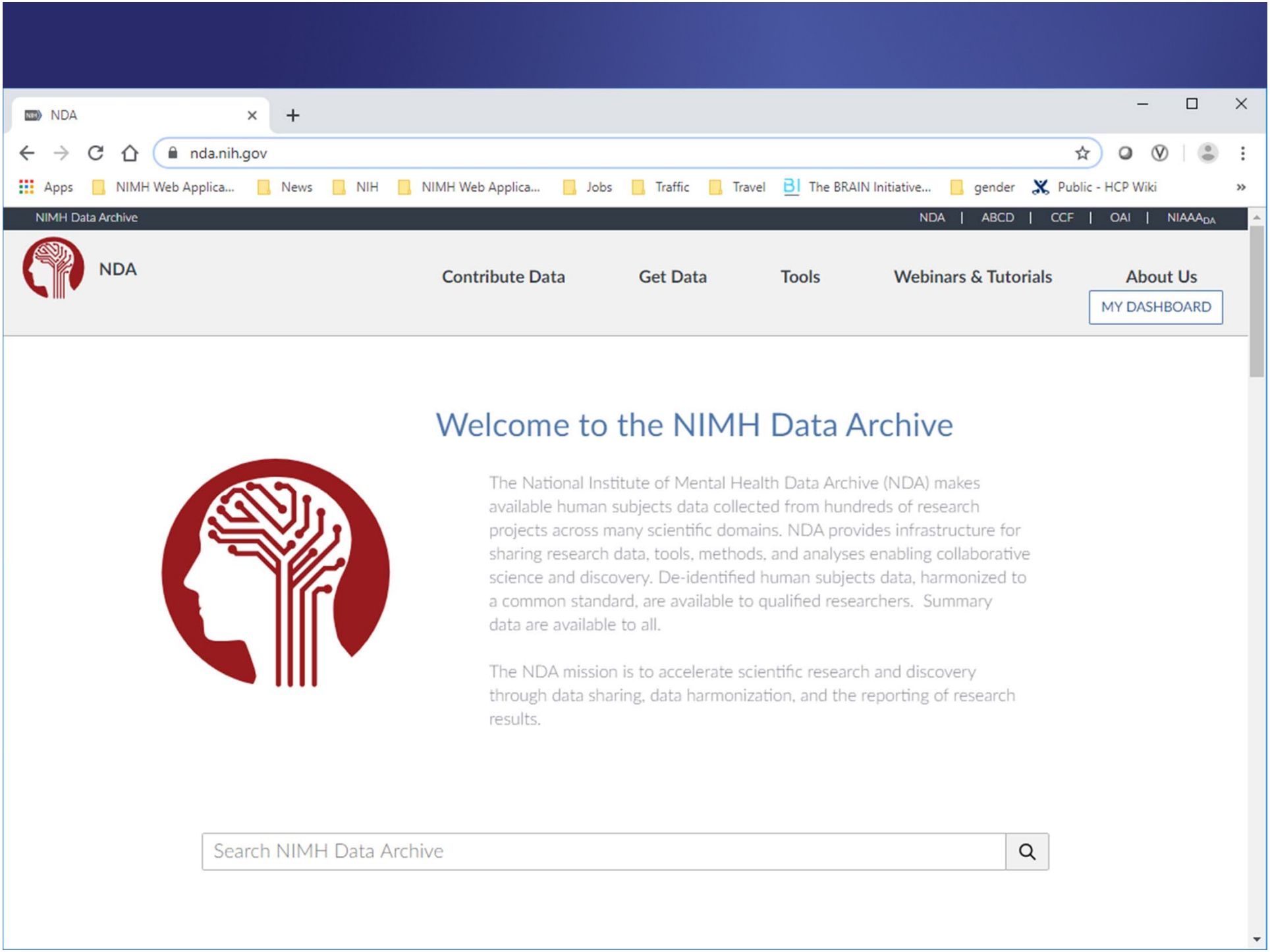
- Responding to a number of instances of high visibility/impact experiments that were not thoughtfully designed, NIH (and NIMH) have instituted a number of programs to enhance rigor and reproducibility in research supported by NIH.
- These discussions with the community started in June 2012. The new guidelines to increase rigor and reproducibility are outlined in NOT-OD-15-103 and at a web site (https://www.nih.gov/research-training/rigor-reproducibility).
- Data archives play an important role in improving the rigor and reproducibility of NIMH funded research through the validation tool (details later)

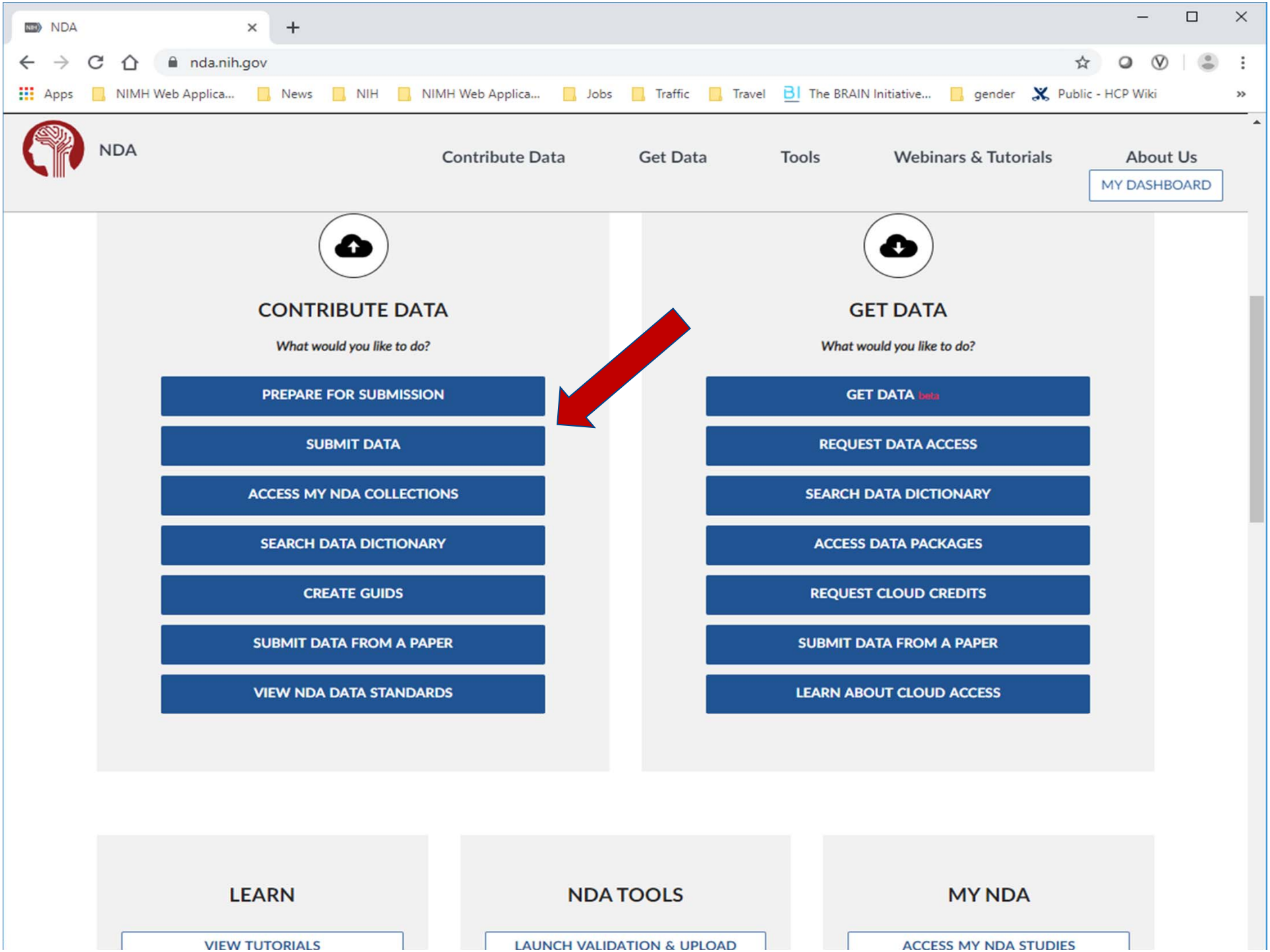
NIMH Data Archive – Current Status

- Holding data from 530,000 research participants.
- Have roughly 2,000 approved users (access lapses after 12 months)
- Storing ~2.0 PB of imaging and –omic data in the Amazon cloud.
- Have 1,000 collections (a data set associated with a grant award)
- Have 210 studies (a data set associated with a publication and assigned a doi)
- (Sadly) have nearly 2,500 different data dictionaries (mainly a clinical data collection instrument like the PHQ-9). We work to harmonize individual items in our data dictionaries, but we currently have nearly 220,000 data elements, so this is a losing battle.



National Institute of Mental Health


NDA
nda.nih.gov
Apps
NIMH Web Applica...
News
NIH
NIMH Web Applica...
Jobs
Traffic
Travel
The BRAIN Initiative...
gender
Public - HCP Wiki
NIMH Data Archive
NDA | ABCD | CCF | OAI | NIAAA DA
NDA
Contribute Data
Get Data
Tools
Webinars & Tutorials
About Us
MY DASHBOARD
Welcome to the NIMH Data Archive
The National Institute of Mental Health Data Archive (NDA) makes available human subjects data collected from hundreds of research projects across many scientific domains. NDA provides infrastructure for sharing research data, tools, methods, and analyses enabling collaborative science and discovery. De-identified human subjects data, harmonized to a common standard, are available to qualified researchers. Summary data are available to all.
The NDA mission is to accelerate scientific research and discovery through data sharing, data harmonization, and the reporting of research results.
Search NIMH Data Archive


NDA
nda.nih.gov
Apps
NIMH Web Applica...
News
NIH
NIMH Web Applica...
Jobs
Traffic
Travel
The BRAIN Initiative...
gender
Public - HCP Wiki
NDA
Contribute Data
Get Data
Tools
Webinars & Tutorials
About Us
MY DASHBOARD
CONTRIBUTE DATA
What would you like to do?
PREPARE FOR SUBMISSION
SUBMIT DATA
ACCESS MY NDA COLLECTIONS
SEARCH DATA DICTIONARY
CREATE GUIDS
SUBMIT DATA FROM A PAPER
VIEW NDA DATA STANDARDS
GET DATA
What would you like to do?
GET DATA beta
REQUEST DATA ACCESS
SEARCH DATA DICTIONARY
ACCESS DATA PACKAGES
REQUEST CLOUD CREDITS
SUBMIT DATA FROM A PAPER
LEARN ABOUT CLOUD ACCESS
LEARN
VIEW TUTORIALS
NDA TOOLS
LAUNCH VALIDATION & UPLOAD
MY NDA
ACCESS MY NDA STUDIES

Prior to Application Submission

1) Use the cost estimation tool to estimate how much it will cost to send the data to NDA. The cost estimation tool does not calculate local data storage or processing costs. It just estimates how much it will cost to work with NDA staff to create new data structures and then to actually send the data to us.
2) Look at the data structures that have already been defined in NDA. Try to reuse existing data structures rather than create new ones. There are useful existing tools to do routine transformations (M/F to 0/1) between local data collection results and the definitions in the data structure.

Internal Work between Submission and Award

1) NDA staff, grants management staff, and program staff work to create the list of all applications recommended for funding that will have to comply with NOT-MH-19-033.

Post Award Expectations

1) PIs can't ignore the NDA Getting Started with Data Sharing notifications that they receive shortly after their Notice of Award has been released. All new awardees get an e-mail letting them know that they have to:
   - Submit a data submission agreement
   - Define the data structures they are going to use to submit data
   - Suggesting (but not requiring) that they use the NDA Global Unique Identifier (GUID) to identify subjects
   - Craft informed consent language that is consistent with sharing broad data sharing with qualified researchers (similar to general research use for genomic data)
2) If an institutional official receives a copy of a letter after the initial award, it means that there is a problem with data sharing that is being escalated.

Post Award Special Cases

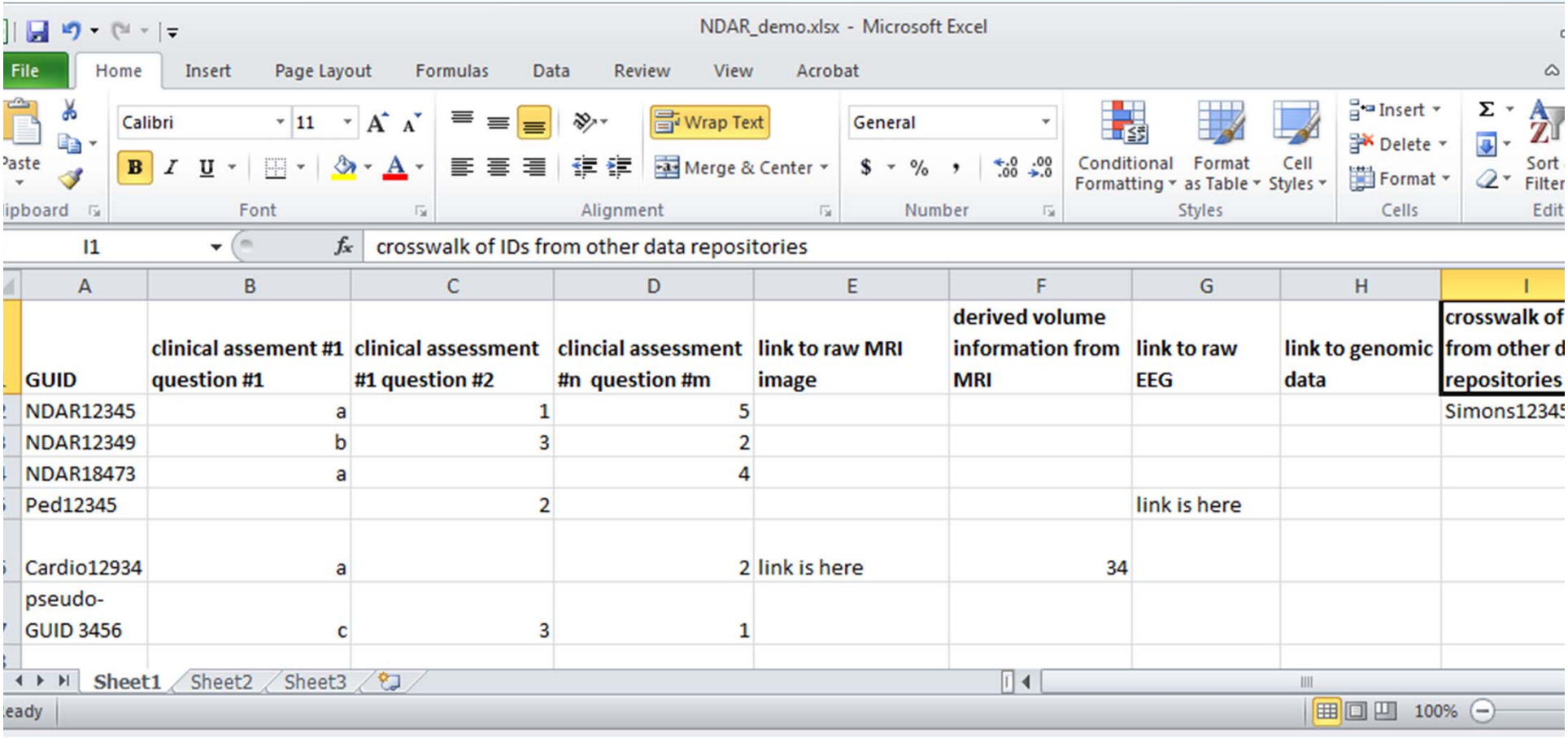
- Consortia – multiple awards
  - There can be two situations here – each awardee is measuring data and there is no central site in the consortium that aggregates the data. In this case, each awardee sends in a data submission agreement and then sends NDA the data
  - The situation where there is a data coordination center that submits the data to NDA on behalf of all of the awardees shouldn't be a significant burden since a single IRB should be governing the consortium
  - There may be special cases. The NDA help desk along with program staff will give guidance.
  - The GUID (or the internal id numbers for the consortia) allow data aggregation.
- Consortia – single award with multiple performance sites
  - The contact institution is responsible for ensuring that data is submitted in accordance with the NDA Terms and Conditions and the data submission agreement.

Other NIH Data Archives

- Many data archives at NIH have data submission processes that are similar to those used at NDA but not identical.
- This could be a good topic for NIH to work with the FDP on harmonizing practices.

NDA Structure – Rows and Columns are the Building Blocks

- It is best to think of NDA as a large (~220,000 data elements by ~530,000 people), sparse, two dimensional matrix.



	A	B	C	D	E	F	G	H	I
	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					

National Institute of Mental Health

Collections and Studies

GUIDS	Data Element 1	Data Element 2	Data Element 3	Data Element 4	Data Element 5	Data Element 6	Data Element 7	Data Element 8	Data Element 9
1									
2									
3									
4									
5									
6									
7									
8									
9									
10									
11									
12									
13									
14									
15									

Data from one collection

Collections and Studies

GUIDS	Data Element 1	Data Element 2	Data Element 3	Data Element 4	Data Element 5	Data Element 6	Data Element 7	Data Element 8	Data Element 9
1									
2									
3									
4									
5									
6									
7									
8									
9									
10									
11									
12									
13									
14									
15									

Data from two collections, no common subjects

Collections and Studies

GUIDS	Data Element 1	Data Element 2	Data Element 3	Data Element 4	Data Element 5	Data Element 6	Data Element 7	Data Element 8	Data Element 9
1									
2									
3									
4									
5									
6									
7									
8									
9									
10									
11									
12									
13									
14									
15									

Data from two collections, no common subjects, but there are common data elements. Harmonization can be “interesting”.

Collections and Studies

GUIDS	Data Element 1	Data Element 2	Data Element 3	Data Element 4	Data Element 5	Data Element 6	Data Element 7	Data Element 8	Data Element 9
1									
2									
3									
4									
5									
6									
7									
8									
9									
10									
11									
12									
13									
14									
15									

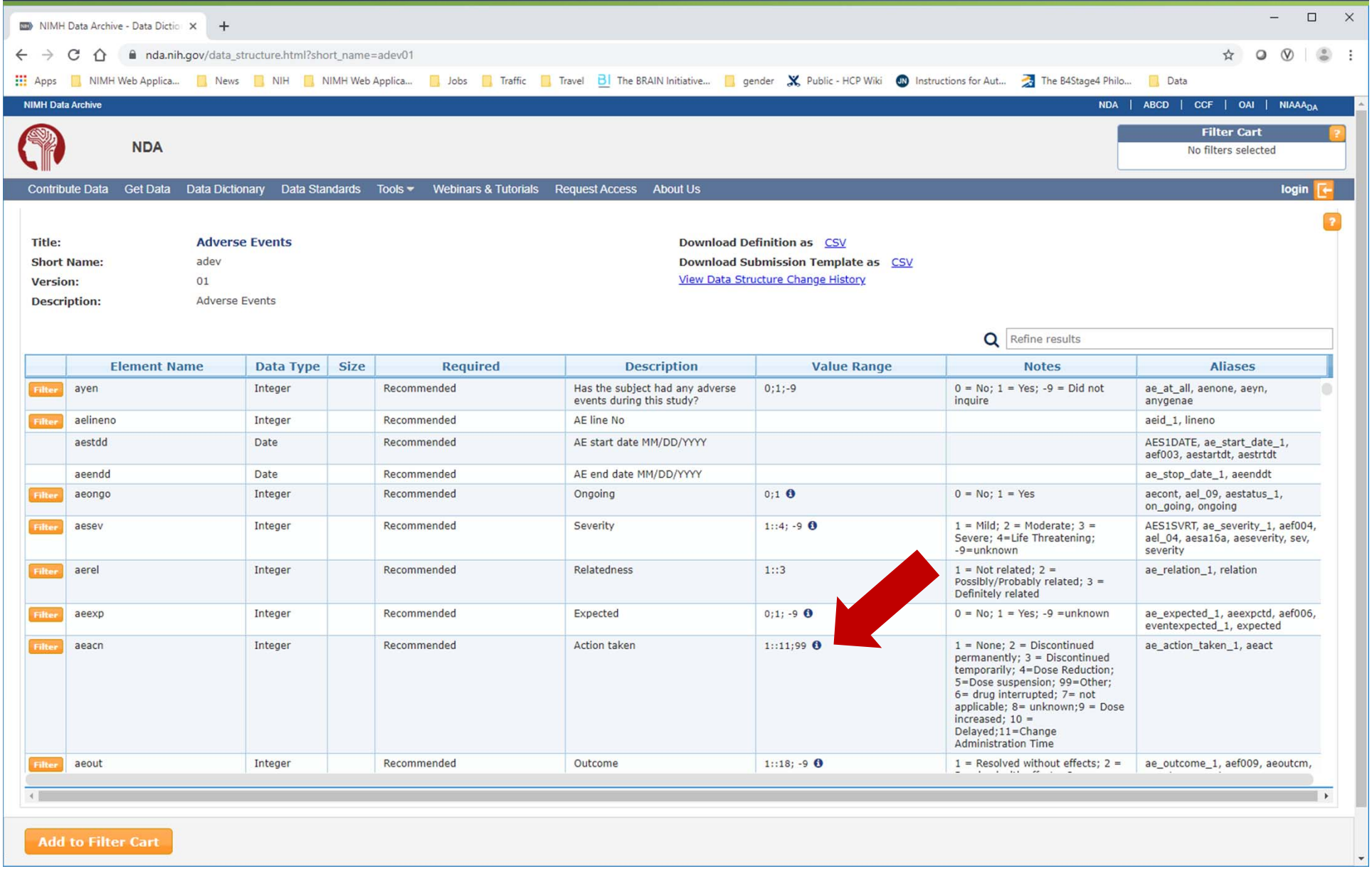
Data from two collections and there are two common subjects (7 and 8). Potentially there are issues here related to the owners of the two collections accessing other data about a single research participant.

Collections and Studies

GUIDS	Data Element 1	Data Element 2	Data Element 3	Data Element 4	Data Element 5	Data Element 6	Data Element 7	Data Element 8	Data Element 9
1									
2									
3									
4									
5									
6									
7									
8									
9									
10									
11									
12									
13									
14									
15									

Data from one collection with additional derived information from a study added

Typical Data Structure



	Element Name	Data Type	Size	Required	Description	Value Range	Notes	Aliases
Filter	ayen	Integer		Recommended	Has the subject had any adverse events during this study?	0;1;-9	0 = No; 1 = Yes; -9 = Did not inquire	ae_at_all, aenone, aeyn, anygenae
Filter	aelineno	Integer		Recommended	AE line No			aeid_1, lineno
	aestdd	Date		Recommended	AE start date MM/DD/YYYY			AES1DATE, ae_start_date_1, aef003, aestartdt, aestrtdt
	aeendd	Date		Recommended	AE end date MM/DD/YYYY			ae_stop_date_1, aeenddt
Filter	aeongo	Integer		Recommended	Ongoing	0;1	0 = No; 1 = Yes	aecont, ael_09, aestatus_1, on_going, ongoing
Filter	aesev	Integer		Recommended	Severity	1::4; -9	1 = Mild; 2 = Moderate; 3 = Severe; 4=Life Threatening; -9=unknown	AES1SVRT, ae_severity_1, aef004, ael_04, aesa16a, aeseverity, sev, severity
Filter	aerel	Integer		Recommended	Relatedness	1::3	1 = Not related; 2 = Possibly/Probably related; 3 = Definitely related	ae_relation_1, relation
Filter	aeexp	Integer		Recommended	Expected	0;1; -9	0 = No; 1 = Yes; -9 =unknown	ae_expected_1, aeexpctd, aef006, eventexpected_1, expected
Filter	aeacn	Integer		Recommended	Action taken	1::11;99	1 = None; 2 = Discontinued permanently; 3 = Discontinued temporarily; 4=Dose Reduction; 5=Dose suspension; 99=Other; 6= drug interrupted; 7= not applicable; 8= unknown;9 = Dose increased; 10 = Delayed;11=Change Administration Time	ae_action_taken_1, aeact
Filter	aeout	Integer		Recommended	Outcome	1::18; -9	1 = Resolved without effects; 2 =	ae_outcome_1, aef009, aeoutcm,

Data Access

- NDA has a data access form that requests similar information to the dbGaP data access form.
- The goal of all NIH data access forms is to ensure that the data are used in ways that are consistent with the informed consents.
- For NDA, we really care about preventing any attempt at re-identification as well as preventing redistribution of the data.
- We would be interested in hearing more about the data use certification work going on at FDP and how we might be able to make use of a global submission agreement

Conclusion

- We are reasonable people trying to find ways to meaningfully aggregate data to allow better understanding of complex and painful diseases.
- We really are interested in working with you to minimize the burdens that come with data aggregation
- We do think that both the secondary data analysis and the improvements in rigor and reproducibility are worth the effort that it takes to submit data to the NIMH Data Archive.