



Integrating Externally Acquired Data Into Standard Workflows

Triumphs and Challenges

Phuse 2018

What is External Data?

- External Data is Data Acquired where we are not the sponsor
- We are not the sponsor... we don't have control over the data collection/protocol.
 - No insight into quality of the data
 - “Completeness” / Accuracy / Ability to trace the data from source to final dataset
 - Usually just get a bunch of data... lucky if you get a data dictionary.

Examples - 1

 USC Mark and Mary Stevens Neuroimaging and Informatics Institute

 USC University of Southern California



SEARCH

[Access Data](#) | [Upload Data](#) | [Contact Us](#)

ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE

ABOUT | STUDY DESIGN | DATA & SAMPLES | METHODS & TOOLS | SUPPORT | NEWS & PUBLICATIONS





Sharing Alzheimer's Research Data with the World

The Alzheimer's Disease Neuroimaging Initiative (ADNI) unites researchers with study data as they work to define the progression of Alzheimer's disease. ADNI researchers collect, validate and utilize data such as MRI and PET images, genetics, cognitive tests, CSF and blood biomarkers as predictors for the disease. Data from the North American ADNI's study participants, including Alzheimer's disease patients, mild cognitive impairment subjects and elderly controls, are available from this site.



Example – AIBL

- We have some control since we purchased it.
 - Partnership / Paid for it.
 - Caveat Emptor – No return policy

Examples - 2

Parkinson's Progression Markers Initiative

Subscribe: News



PARKINSON'S
PROGRESSION
MARKERS
INITIATIVE

Play a Part in Parkinson's Research

SEARCH

[Contact Us](#) 
[Data Login](#) 
[WG Login](#) 

[Home](#) [About PPMI](#) [Study Design](#) [Access Data & Specimens](#) [Publications & Presentations](#) [PPMI News](#)

OUR MISSION

The mission of PPMI is to identify one or more biomarkers of Parkinson's disease progression. The discovery of a biomarker is a critical step in the development of new and better treatments for PD. This study is being sponsored by **The Michael J. Fox Foundation for Parkinson's Research.**

Subscribe to PPMI e-news



PPMI ENROLLMENT STATUS

A LANDMARK STUDY OF PARKINSON'S DISEASE

The Parkinson's Progression Markers Initiative (PPMI) is a landmark observational clinical study to comprehensively evaluate **cohorts of significant interest** using advanced imaging, biologic sampling and clinical and behavioral assessments to identify biomarkers of Parkinson's disease progression.

PPMI is taking place at **clinical sites** in the United States, Europe, Israel, and Australia. Data and samples acquired from study participants will enable the development of a comprehensive Parkinson's **database and biorepository**, which is currently available to the scientific community to conduct field-changing research.

PPMI is made possible by the concerted efforts of a number of collaborators. This study is sponsored by The Michael J. Fox Foundation for Parkinson's Research.

Learn more about **Who We Are**.

LATEST NEWS FROM PPMI

PPMI Study Update Call – 3/15/18
PPMI Study Update Call Thursday, March 15 12-1 pm Eastern Time Andrew Siderowf, MD, MSCE Professor of Neurology University of Pennsylvania Co-Principal Investigator of

 **DOWNLOAD DATA**

 **REQUEST SPECIMENS**

 **REQUEST CELL LINES**

 **for PROSPECTIVE PARTICIPANTS**

How do we get this data?



Step 1 in the Workflow: Acquire the Data

- Easier Said Than Done
 - Download from a Website?
 - Email
 - Not in a Database (csv)

But I'm Sure It's In A Good Format

Problems:

- Do you like CSVs?
- Like, a lot of CSV's (>100)

Name	Date modified	Type	Size
 TOMM40	1/28/2016 11:19 A...	Microsoft Excel Co...	40 KB
 TREATDIS	1/23/2017 7:12 AM	Microsoft Excel Co...	322 KB
 UASPMVBM	1/23/2017 8:05 AM	Microsoft Excel Co...	5,454 KB
 UCBERKELEYAV45_10_17_16	1/23/2017 8:06 AM	Microsoft Excel Co...	4,039 KB
 UCBERKELEYAV1451_10_17_16	1/23/2017 8:06 AM	Microsoft Excel Co...	190 KB
 UCBERKELEYAV1451_PVC	1/23/2017 8:06 AM	Microsoft Excel Co...	141 KB
 UCBERKELEYFDG_DICT_03_11_13	1/23/2017 8:06 AM	Microsoft Excel Co...	5 KB
 UCBERKELEYFDG_07_30_15	1/23/2017 8:06 AM	Microsoft Excel Co...	2,517 KB
 UCD_ADNI1_WMH	1/23/2017 8:05 AM	Microsoft Excel Co...	380 KB
 UCD_ADNI2_WMH_10_26_15	1/23/2017 8:05 AM	Microsoft Excel Co...	498 KB
 UCSDVOL	1/23/2017 8:05 AM	Microsoft Excel Co...	711 KB
 UCSFASLFS_11_02_15_V2	1/23/2017 8:05 AM	Microsoft Excel Co...	3,662 KB
 UCSFASLQC	1/23/2017 8:06 AM	Microsoft Excel Co...	123 KB
 UCSFATRPHY	1/23/2017 8:06 AM	Microsoft Excel Co...	185 KB
 UCSFFSL_02_01_16	1/23/2017 8:06 AM	Microsoft Excel Co...	9,361 KB
 UCSFFSL51ALL_08_01_16	1/23/2017 8:06 AM	Microsoft Excel Co...	5,359 KB
 UCSFFSL51Y1_08_01_16	1/23/2017 8:06 AM	Microsoft Excel Co...	3,658 KB
 UCSFFSX_11_02_15	1/23/2017 8:06 AM	Microsoft Excel Co...	11,601 KB
 UCSFFSX51_08_01_16	1/23/2017 8:05 AM	Microsoft Excel Co...	12,384 KB
 UCSFFSX51_11_02_15_V2	1/23/2017 8:05 AM	Microsoft Excel Co...	12,061 KB
 UCSFFSX51_ADNI1_3T_02_01_16	1/23/2017 8:05 AM	Microsoft Excel Co...	1,345 KB
 UCSFSNTVOL	1/23/2017 8:06 AM	Microsoft Excel Co...	266 KB
 UPENNBIOMK	1/23/2017 7:18 AM	Microsoft Excel Co...	37 KB
 UPENNBIOMK_MASTER	1/23/2017 7:18 AM	Microsoft Excel Co...	710 KB
 UPENNBIOMK2	1/23/2017 7:18 AM	Microsoft Excel Co...	57 KB

But I'm Sure It's Well-Defined

Problems:

- Do you like 8 character alpha-numeric variable names?
- Like, a lot of 8 character alpha-numeric variable names (> 100)?

64	ST110CV	Num	8 BEST5.	BEST5.
65	ST110SA	Num	8 BEST4.	BEST4.
66	ST110TA	Num	8 BEST5.	BEST5.
67	ST110TS	Num	8 BEST5.	BEST5.
68	ST111CV	Num	8 BEST5.	BEST5.
69	ST111SA	Num	8 BEST4.	BEST4.
70	ST111TA	Num	8 BEST5.	BEST5.
71	ST111TS	Num	8 BEST5.	BEST5.
72	ST112SV	Num	8 BEST4.	BEST4.
73	ST113CV	Num	8 BEST4.	BEST4.
74	ST113SA	Num	8 BEST4.	BEST4.
75	ST113TA	Num	8 BEST5.	BEST5.
76	ST113TS	Num	8 BEST5.	BEST5.
77	ST114CV	Num	8 BEST5.	BEST5.
78	ST114SA	Num	8 BEST4.	BEST4.
79	ST114TA	Num	8 BEST5.	BEST5.
80	ST114TS	Num	8 BEST5.	BEST5.
81	ST115CV	Num	8 BEST5.	BEST5.
82	ST115SA	Num	8 BEST5.	BEST5.
83	ST115TA	Num	8 BEST5.	BEST5.
84	ST115TS	Num	8 BEST5.	BEST5.
85	ST116CV	Num	8 BEST5.	BEST5.
86	ST116SA	Num	8 BEST4.	BEST4.
87	ST116TA	Num	8 BEST5.	BEST5.
88	ST116TS	Num	8 BEST5.	BEST5.
89	ST117CV	Num	8 BEST5.	BEST5.
90	ST117SA	Num	8 BEST4.	BEST4.
91	ST117TA	Num	8 BEST5.	BEST5.
92	ST117TS	Num	8 BEST5.	BEST5.
93	ST118CV	Num	8 BEST5.	BEST5.
94	ST118SA	Num	8 BEST4.	BEST4.
95	ST118TA	Num	8 BEST5.	BEST5.
96	ST118TS	Num	8 BEST5.	BEST5.
97	ST119CV	Num	8 BEST4.	BEST4.
98	ST119SA	Num	8 BEST3.	BEST3.
99	ST119TA	Num	8 BEST5.	BEST5.
100	ST119TS	Num	8 BEST5.	BEST5.
101	ST11SV	Num	8 BEST3.	BEST3.
102	ST120SV	Num	8 BEST5.	BEST5.
103	ST121CV	Num	8 BEST4.	BEST4.

So what principals do we want to follow – The Laws of Public Data

- Has to be in a consistent format
- Has to be able to pooled with other public data sets
- Has to be able to pooled with clinical data.



Workflow Part 2: QC The Data

- Compare like against like
 - For example, if you have a 72 month data cut, and then a 90 month data cut, examine the 72 month data in the 90 month data cut.
 - You'd be surprised at what changes!
 - Often a changelog is not provided



Workflow Part 3: Transform The Data

- Transform to SDTM
 - Clinical data is already in this format
 - People will want to pool it with their own trial data
 - You don't have to spend time creating your own standards



Conversion

- Time consuming and difficult
 - No “magic bullet”
 - Every data set defines something like “visit” different... try the best you can.
 - Only take what you need for now.
 - Start with DM, always. 1 Record per patient, good way to start looking at the data
 - Good task for interns, contractors, sworn enemies.....

Analysis

- Remember that this data does not have the cleanliness of clinical data.
 - **Therefore, always check the integrity of your data.**
- We live in a world where Data Management is not especially rigorous...



How do we let researchers know what is available?

- Once people know that there is data available, they will “swarm” to you.
- You want a way to give people access to the data without being bogged down with hundreds of shells.



Public Data Dashboard

- Goals of the Data Dashboard:
 - View Data
 - Details
 - Don't let them extract it though... it's a guide. They need to talk to a stat first.



How to Program It

- Lots of options, but R/Shiny was the most robust.
- Quick to iterate and get up to speed, but custom enough to allow custom work.



Conclusions

- With a small amount of upfront work, you can enhance the work done by so many departments.
 - HEOR (Health Effectiveness/Outcome Research)
 - Epidemiology
 - Stat Genetics
 - Biomarkers/Imaging
 - And So Much More!