

Integrating Externally Acquired Data Into Standard Workflows

Saranya Duraisamy, Biogen, Cambridge, USA

Nate Mockler, Biogen, Cambridge, USA

ABSTRACT

Externally Acquired Data is defined as data acquired external to clinical development in support of early development research projects. Some examples of this data is the Australian Imaging, Biomarker & Lifestyle Flagship Study of Ageing (AIBL) and Parkinson's Progression Markers Initiative (PPMI). It is sourced from academic collaborations, disease foundations, and purchases, and contains information including demographic surveys, questionnaires, imaging, lab, and CSF data. This can be challenging to integrate it into standard programming practices, as they may be in different formats, structures, and have varying levels of quality. This talk will describe the process of integrating, transforming, analyzing and creating dashboards to visualize this data.

INTRODUCTION

Your company just made a huge announcement: they've purchased an external data repository. Or maybe they've entered a consortium with some other companies. Perhaps your company has entered into an arrangement with a country's health database, an insurance company, or electronic medical repositories. You were not privy to this decision; you received the email like everyone else. Then you start receiving the emails: statisticians want to do analyses on this data, others want to integrate with your clinical trial data. Marketing and other internal groups want their own copies and not tell you what they're doing with. Oh, and you're in charge of keeping track of it all. Uh oh.

EXTERNAL AFFAIRS

How do we define external data? Definitions can vary, but the one that I always liked is that external Data is Data Acquired where we are not the sponsor. This is not referring to sales data, but clinical data that brought internally to supplement existing analyses or serve a precursor to a future analysis. Since your company is not the sponsor, they have no control over the data. This also means there is no control of the data collection process or protocol. A data dictionary may be included, it may not be. That means it may be challenging to trace the data from the CRF to the final state.

One example is the Alzheimer's Disease NeuroImaging Initiative (ADNI). ADNI is a collaboration between numerous pharmaceutical companies and the National Institute on Aging. It contains longitude data in such areas as clinical data, imaging (PET and MRI data), along with select genetic and biospecimen data. A company can simply access this data rather than run their own study at a significant cost of time and money. It has also led to offshoots such as the Department of Defense ADNI, and the Australian Imaging, Biomarker & Lifestyle Flagship Study of Ageing (AIBL). Especially in this era of "big data" and analytics that require large pools of data, the ability to plug in a large amount of both normal and affected patients over long period of time is a great asset.

STEP 1: ACQUIRING THE DATA

Now let's look at one of the packages that are in the "raw data" section of the download that you made receive:

PhUSE US Connect 2018

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
 UCSFATRPY	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
 UPENNBBIOMK	1/23/2017 7:18 AM	Microsoft Excel Co...	37 KB
 UPENNBBIOMK_MASTER	1/23/2017 7:18 AM	Microsoft Excel Co...	710 KB
 UPENNBBIOMK2	1/23/2017 7:18 AM	Microsoft Excel Co...	57 KB

Output 1: That's a Lot of Files

That can be a challenge. What can be more challenging is if the variables are not labeled very clearly...

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.

Output 2: That's a Lot of Variables

PhUSE US Connect 2018

To deal with these issues, there are three laws of Public Data that I tend to stick with:

- 1) The data has to be a consistent format
- 2) The data has to be able to be pooled with other public data sets
- 3) The data has to be able to be pooled with other clinical data sets.

Let's talk about that. For the data being a consistent format, for analyses sake. Also, the data must be able to be pooled for other public data sets. Your investigators will probably end up pooling it anyway with the other public data sets, so might as well do the work beforehand. Same with pooling it with other clinical data sets... your people are going to do it anyway, might as well make it easier, and therefore let's make it SDTM.

For acquisition of the data, first the process has to be defined? Who is going to be downloading the data? How often will the data be updated? That has to be well documented.

STEP 2: QCING THE DATA

For QCing the data, often the data is not double programmed or audited for a rigorous validation process. The first step is to compare like against like. When looking at say a 12 month data vs a 6 month data set, you want to compare the 6 month data with the 6 month data in the 12 month data. It can be surprising what data can change.... There may not be a changelog for this.

STEP 3: CONVERT THE DATA

The best way to handle this to convert the public data to SDTM. Now, that's a lot of work, but rather than have a bunch of groups do it piecemeal, doing it all at once leads to a much more efficient process. Also, it may not have to be done all at once... maybe start with DM and LB, for instance.

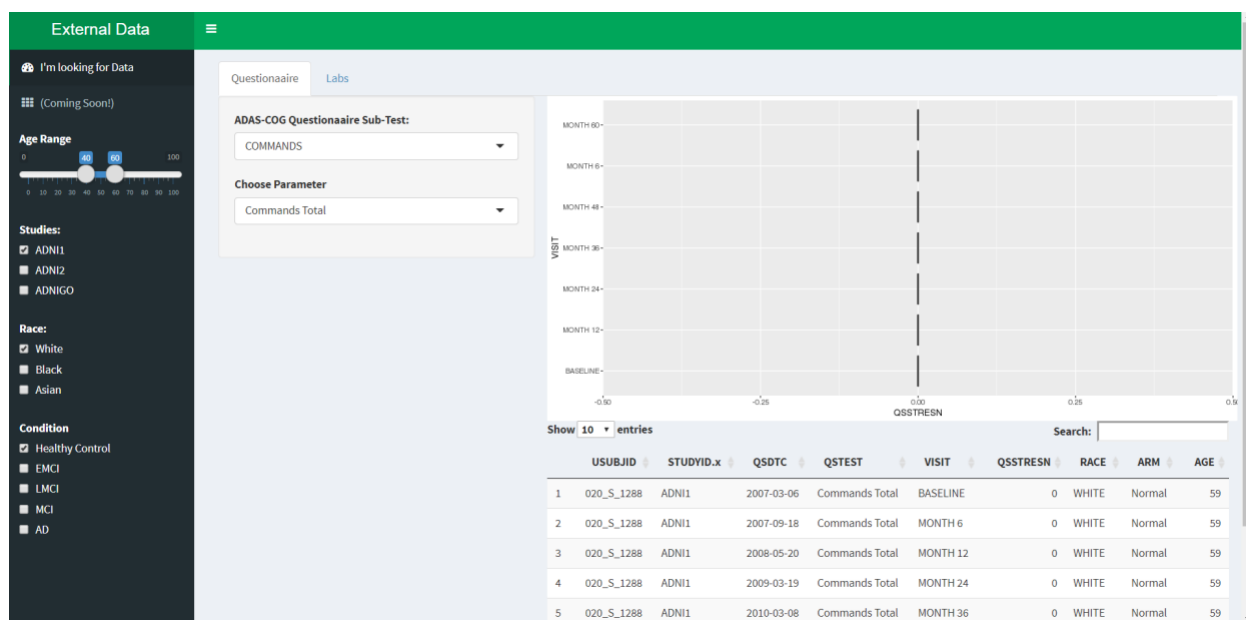
ANALYSIS

Make sure to do a lot of data cleaning. It is critical to understand that the data cleaning process may not be as rigorous as is done otherwise.

PUBLIC DATA DASHBOARD

So now your data is analyzed, transformed, and ready to go. How do you make sure that people know where the data is and what kind of tests are available? The best way to do so is through a dashboard... either Tableau, SAS VA, or R/Shiny for instance? For this example, we'll talk about R/Shiny, but there are a lot of other great ways to do so.

PhUSE US Connect 2018



Output 3: Letting your users do the work so you don't have to

An example of the dashboard is shown above. There are two parts in the dashboard: "I'm Looking For Data", and "Individual Patient Explorer". The user can select the following filters: age range, studies, race, and medical condition. Once the population is selected, the user can pick a data set (such as QS or LB) and a certain test and subtest. Once that happens, box plots and tables can be displayed and downloaded. One can then go to Individual Patient Explorer, and download an individual patient profile.

CONCLUSION

With a small amount of work up-front, you can enhance the work done by so many departments. Groups such as Epidemiology, Health Economy, and Statistical Genetics can access the large amount of data and assist in their work, helping to lead to scientific breakthroughs. Also, it can help your job as well... letting them handle the basic analyses so you can focus on high-value work.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Author Name: Nate Mockler and Saranya Duraisamy

Company: Biogen

Email: Nate.Mockler@biogen.com & Saranya.Duraisamy@biogen.com

Web: www.biogen.com

Brand and product names are trademarks of their respective companies.