Research that you and others can reuse & reproduce

Day 7 | Session 3
July 13, 2021
KCNI Summer School 2021

Abhi Pratap, PhD
Group Head - Digital Health & AI
Independent Scientist, KCNI
Faculty Affiliate, Vector Institute

@abhishekpratap

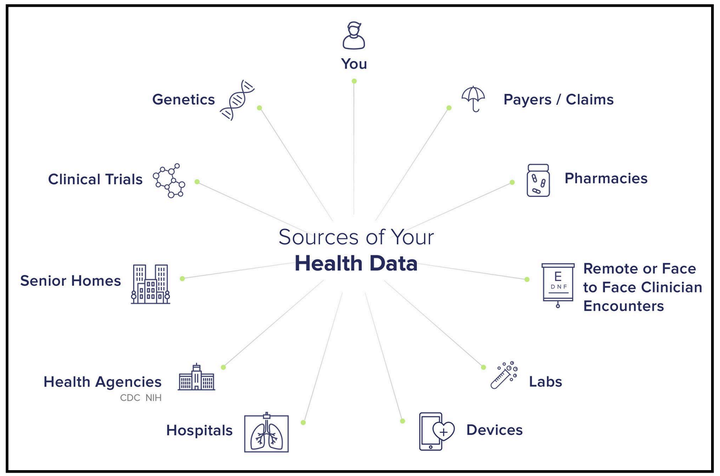
camh Centre for Addiction and Mental Health | Krembil Center for Neuroinformatics

Housekeeping - In case you want to follow along

1. Create an account on synapse - www.syanpse.org

2. If you want to upload data - Get certified - https://www.synapse.org/#!Quiz:Certification

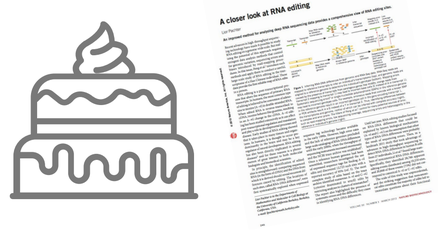
“Context”



Sources **Data** **Insights**

Taken from - “https://insuranceblog.accenture.com/underwriting-climb-mountain-of-data-in-digital-age"

Recipe | Ingredients



Lab 1 - Original Student

Lab 1 - Second Grad Student

Lab 2

Missing ingredients

Lab 3

Lab 4

and the journal
is is slower
papers.

2012
TOTAL PAPERS PUBLISHED WORLDWIDE
2,121,740

1997
1,143,419

Increase in number of papers submitted to *Nature*
43%

Increase in number of papers published worldwide
86%

1997 → 2012

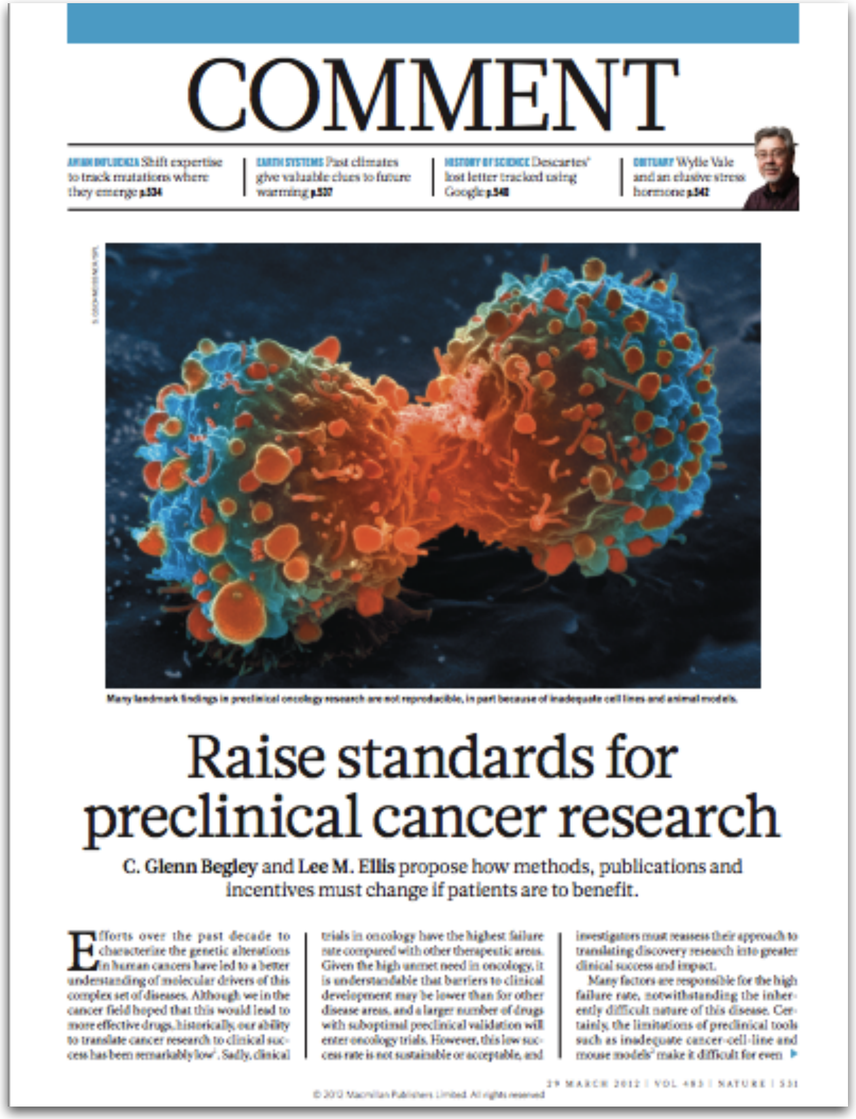
Science publishing: The golden club

mostly not useful

85%

Chalmers and Glasziou 2009

Science publishing: The golden club



COMMENT

AVIAN INFLUENZA Shift expertise to track mutations where they emerge p.534 | EARTH SYSTEMS Past climates give valuable clues to future warming p.537 | HISTORY OF SCIENCE Descartes' lost letter tracked using Google p.548 | OBITUARY Wylie Vale and an elusive stress hormone p.542

Many landmark findings in preclinical oncology research are not reproducible, in part because of inadequate cell lines and animal models.

Raise standards for preclinical cancer research

C. Glenn Begley and **Lee M. Ellis** propose how methods, publications and incentives must change if patients are to benefit.

Efforts over the past decade to characterize the genetic alterations in human cancers have led to a better understanding of molecular drivers of this complex set of diseases. Although we in the cancer field hoped that this would lead to more effective drugs, historically, our ability to translate cancer research to clinical success has been remarkably low[1]. Sadly, clinical trials in oncology have the highest failure rate compared with other therapeutic areas. Given the high unmet need in oncology, it is understandable that barriers to clinical development may be lower than for other disease areas, and a larger number of drugs with suboptimal preclinical validation will enter oncology trials. However, this low success rate is not sustainable or acceptable, and investigators must reassess their approach to translating discovery research into greater clinical success and impact.

Many factors are responsible for the high failure rate, notwithstanding the inherently difficult nature of this disease. Certainly, the limitations of preclinical tools such as inadequate cancer-cell-line and mouse models[2] make it difficult for even ▸

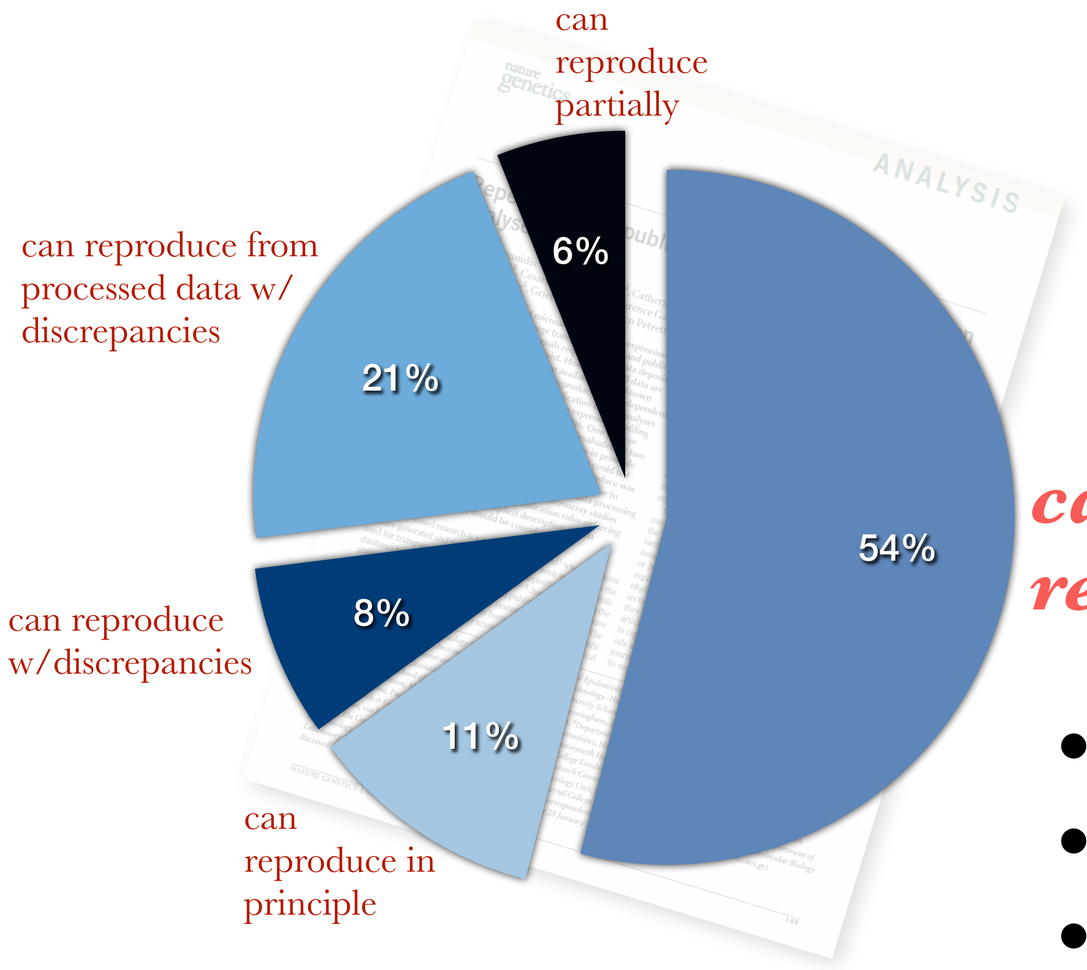
29 MARCH 2012 | VOL 483 | NATURE | 531

Begley & Ellis. Nature (2012)

"The biotechnology firm Amgen...tried to confirm published findings of 53 'landmark papers' … scientific findings were confirmed in only 6 (11%) cases."

"Bayer HealthCare in Germany last year reported that only about 25% of published preclinical studies could be validated to the point at which projects could continue."

The status quo tolerates poor communication of findings



cannot reproduce

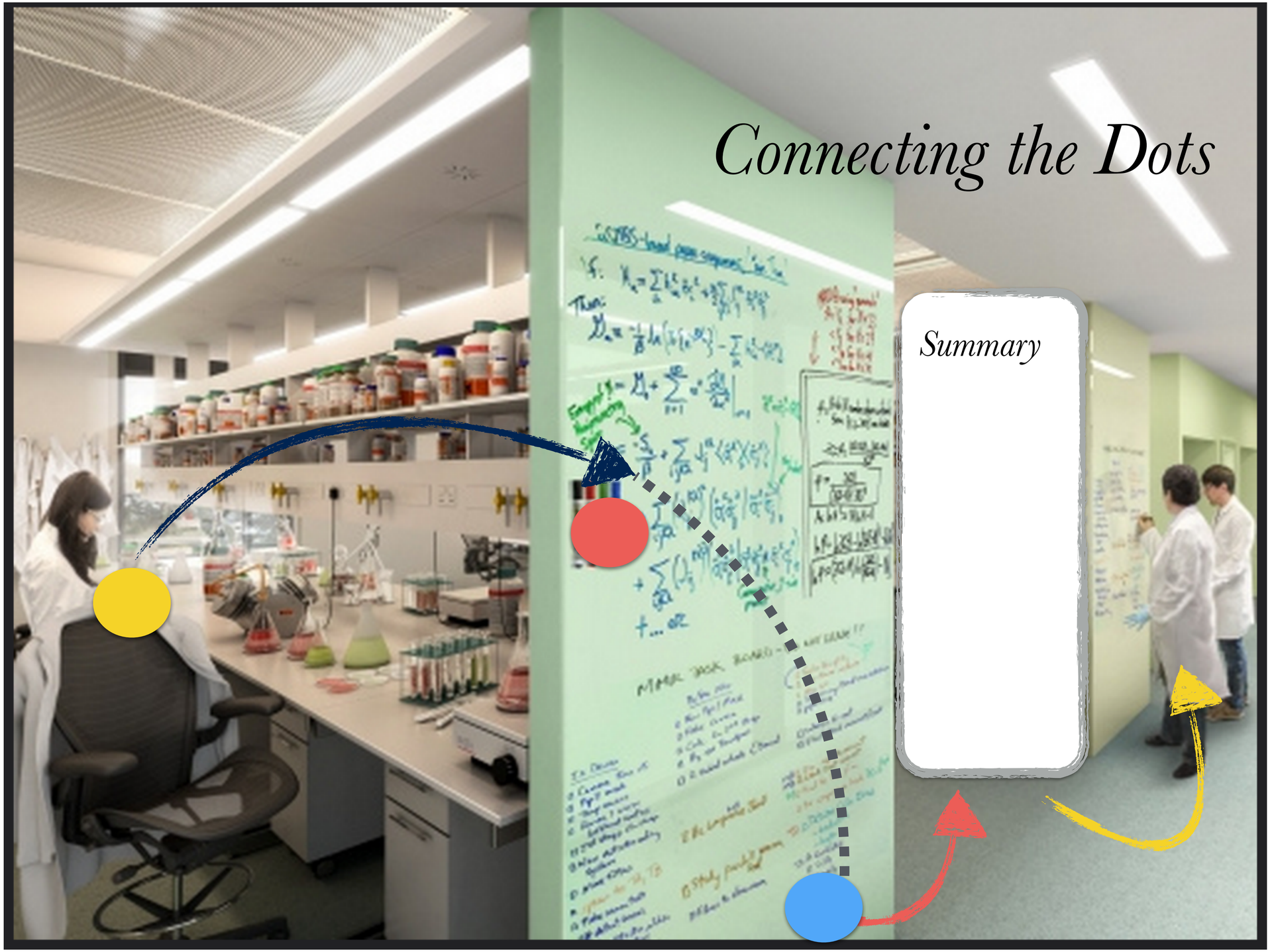
- **Data not available**
- **Software not available**
- **Methods unclear**
- **Different result altogether**

Ioannidis A. et al. *Repeatability of published microarray gene expression analyses*. Nature Genetics 41, 149-155 (2009) | doi:10.1038/ng.295

Source

Mixing ratio

Availability


Connecting the Dots
Summary

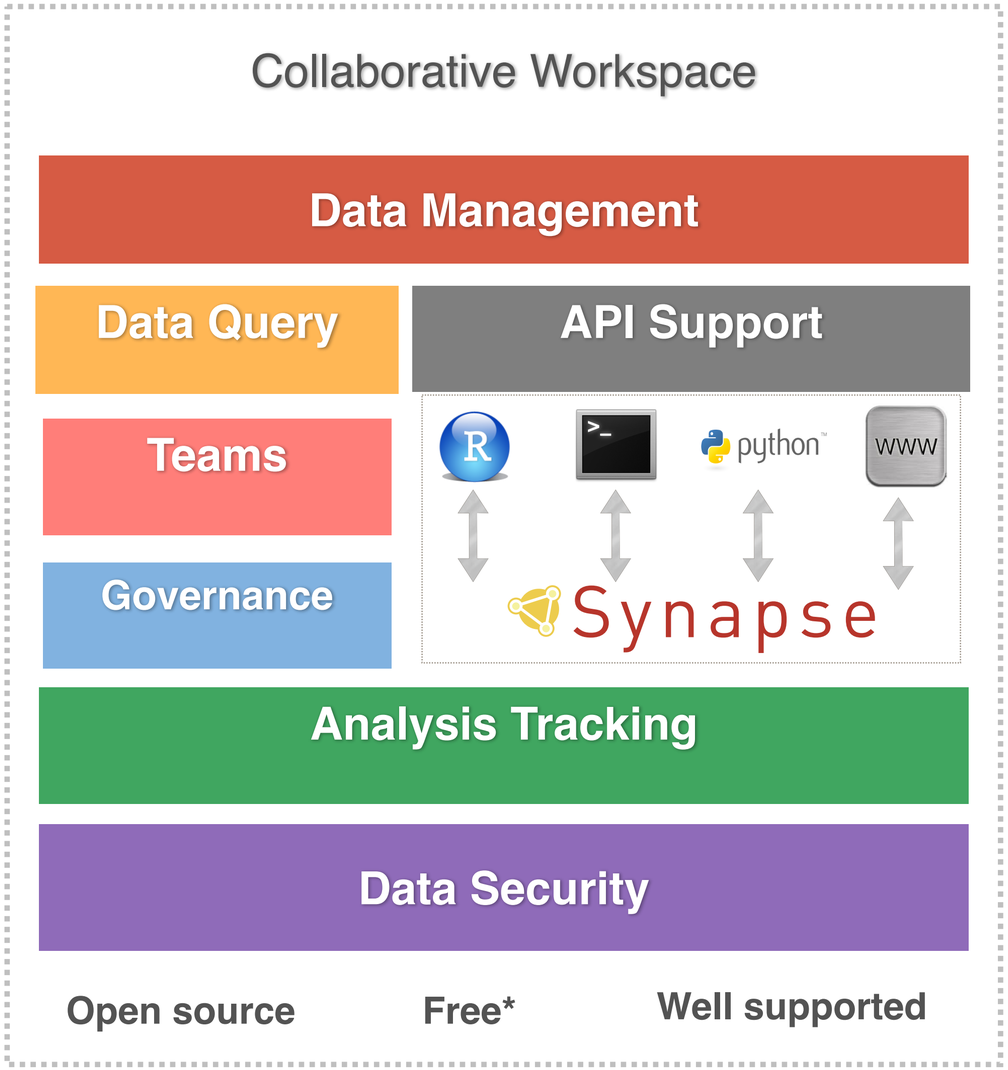
Data

Creation → Curation → Consumption

www.synapse.org

a tool to improve **transparency** and **reproducibility** of data intensive science by recording analyses in **real-time**

a collection of living research projects **enabling** researchers to contribute to large-scale **collaborative science**



PARTNERS

NIH National Heart, Lung, and Blood Institute

NIH National Institute of Mental Health

Alfred P. Sloan FOUNDATION

NIH National Institute on Aging

Life Sciences DISCOVERY FUND

CHILDREN'S TUMOR FOUNDATION ENDING NF THROUGH RESEARCH

NTAP

Many alternatives out there

.....and more

clinical

omics

mobile apps

sensors

surveys

wearables

Creation

various options for Data I/O

Realtime data query & sharing
Complete data provenance

Curation

Resarch Study Portal

forum

Responsible Data Sharing

Governance

Data Annotation

Provenance Tracking

Data Queries

Data Security

Data Versioning

Consumption

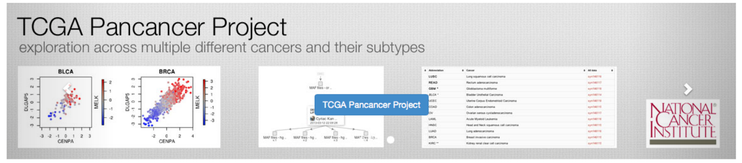
clinical

omics

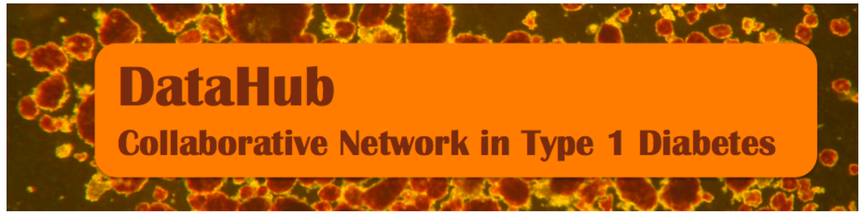
mobile apps

sensors surveys wearables

Creation



International Cancer Genome Consortium



(Omics, Clinical data + Model to data approaches)

Parkinsons

Cardiovascular

MoleMapper

Asthma

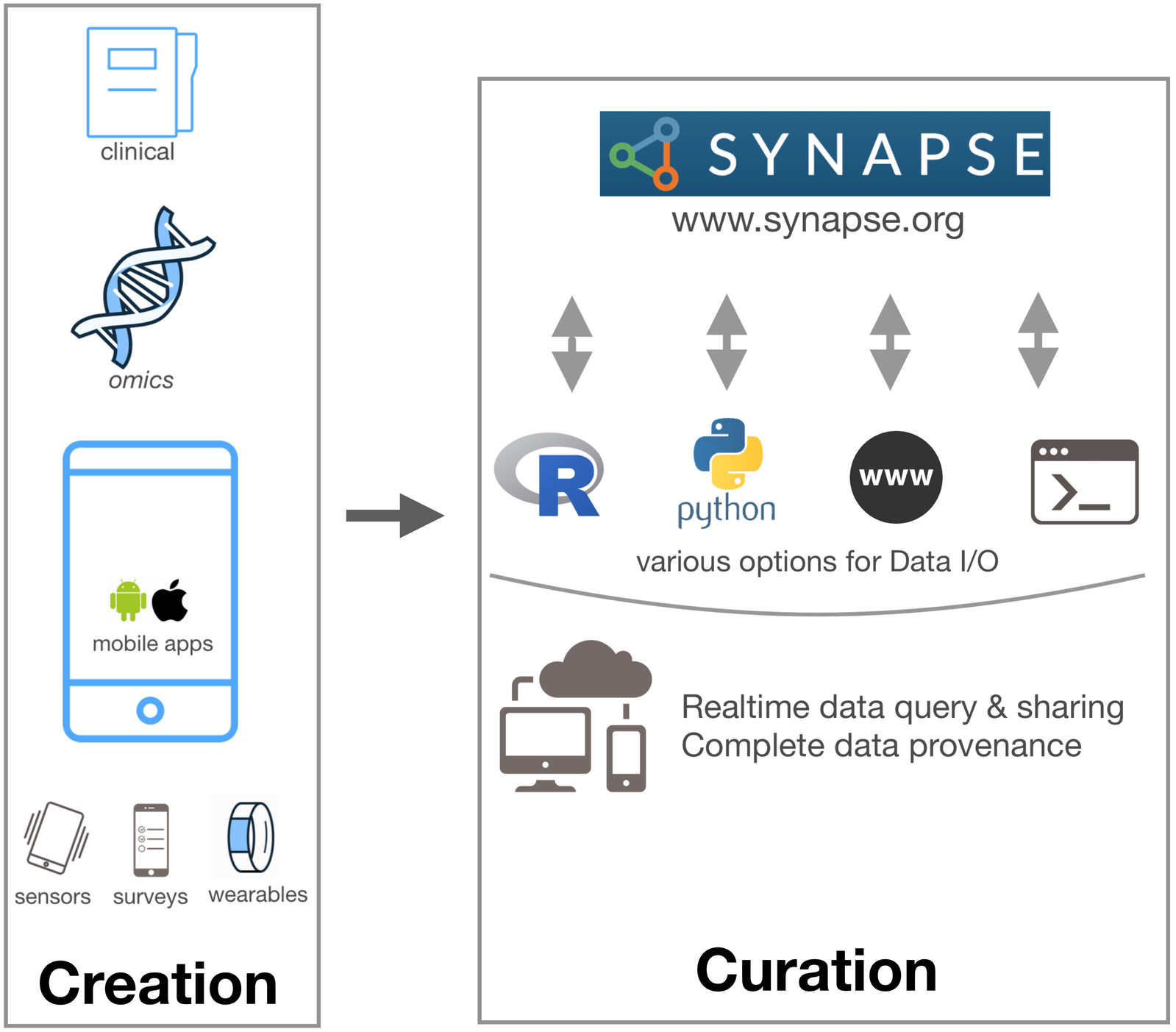
Multiple Sclerosis

AllofUS CRF

Samsung my BP lab

www.synapse.org/digitalhealth

(PROs, sensor-based functional tests, wearables, passive data etc)


clinical
omics
mobile apps
sensors
surveys
wearables
Creation
SYNAPSE
www.synapse.org
python
www
various options for Data I/O
Realtime data query & sharing
Complete data provenance
Curation

SYNAPSE
www.synapse.org

various options for Data I/O

Realtime data query & sharing
Complete data provenance

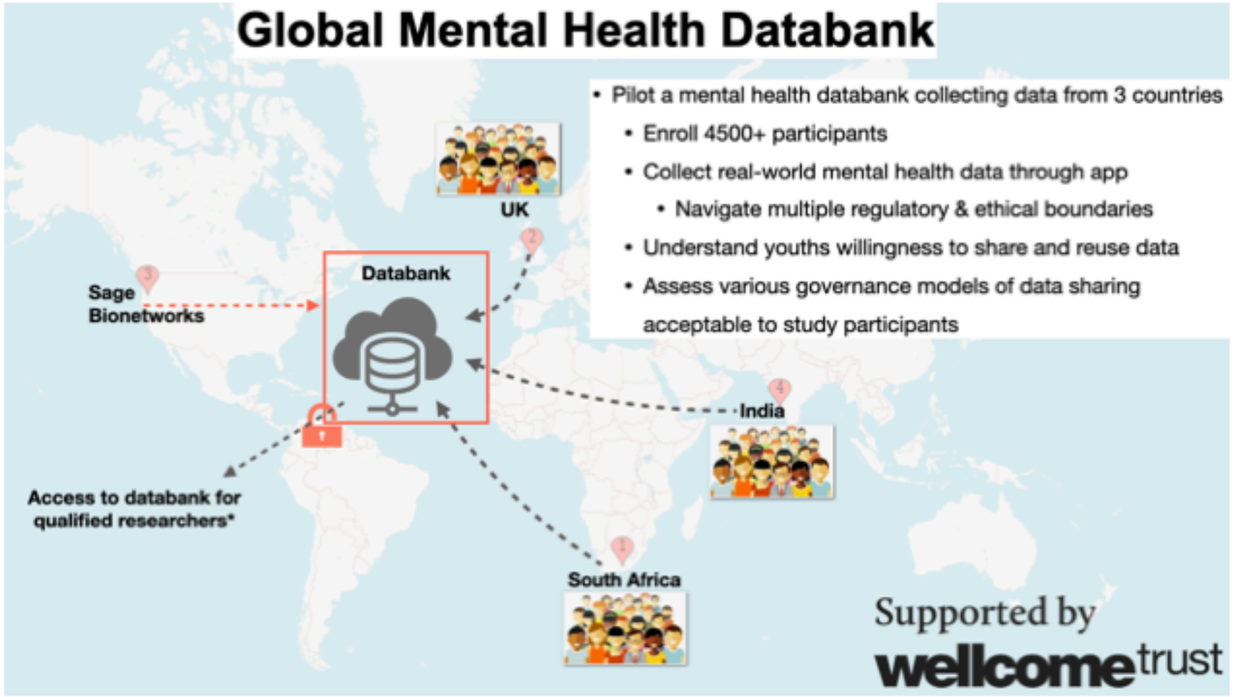
Curation

Wellcome GMHD

UniqueID - syn123456

Project

Summary



Data

Folders

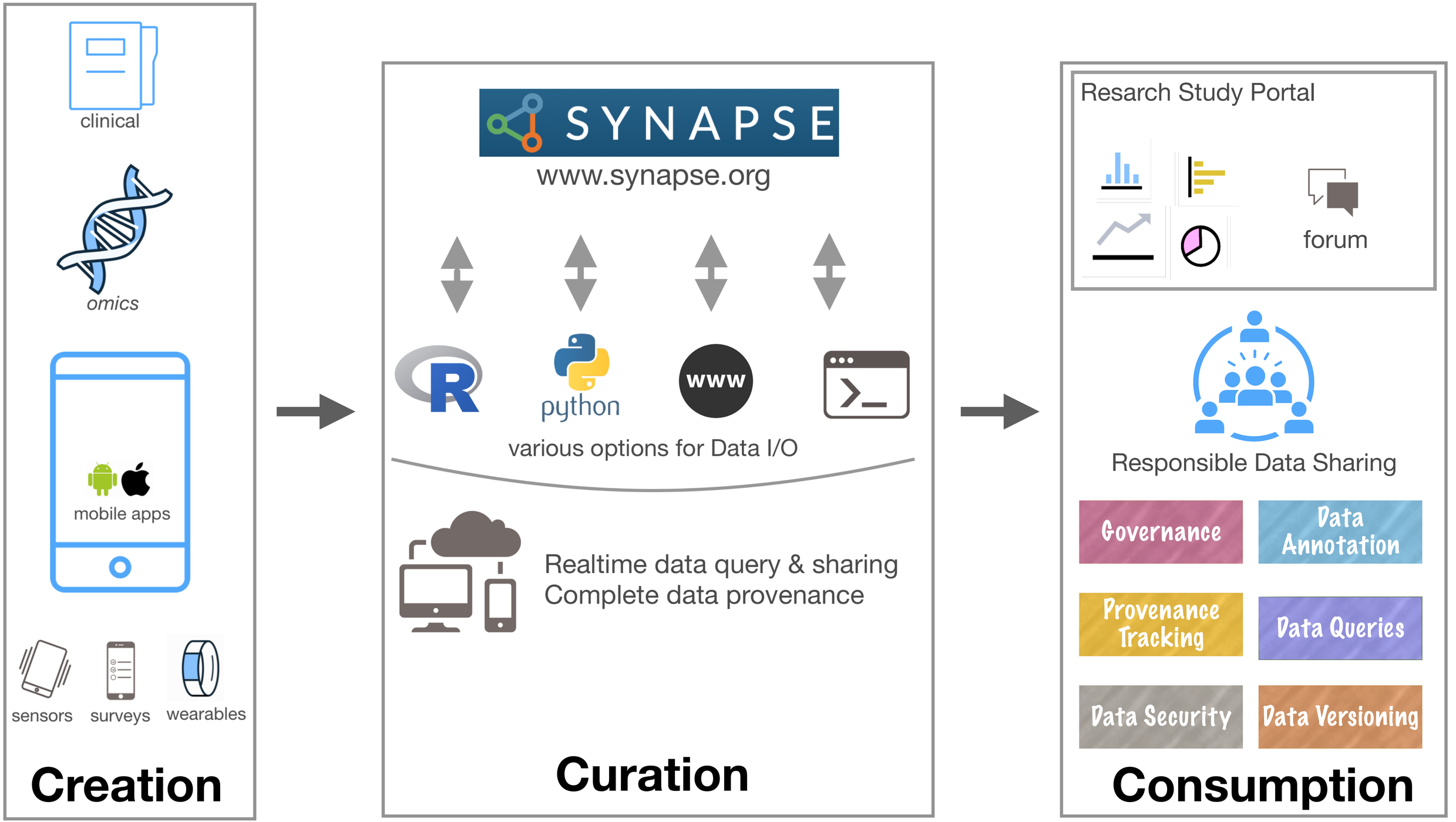
raw data	syn48788
results	syn4785

Files/Tables

PHQ-9	syn85656
Demog	syn47858

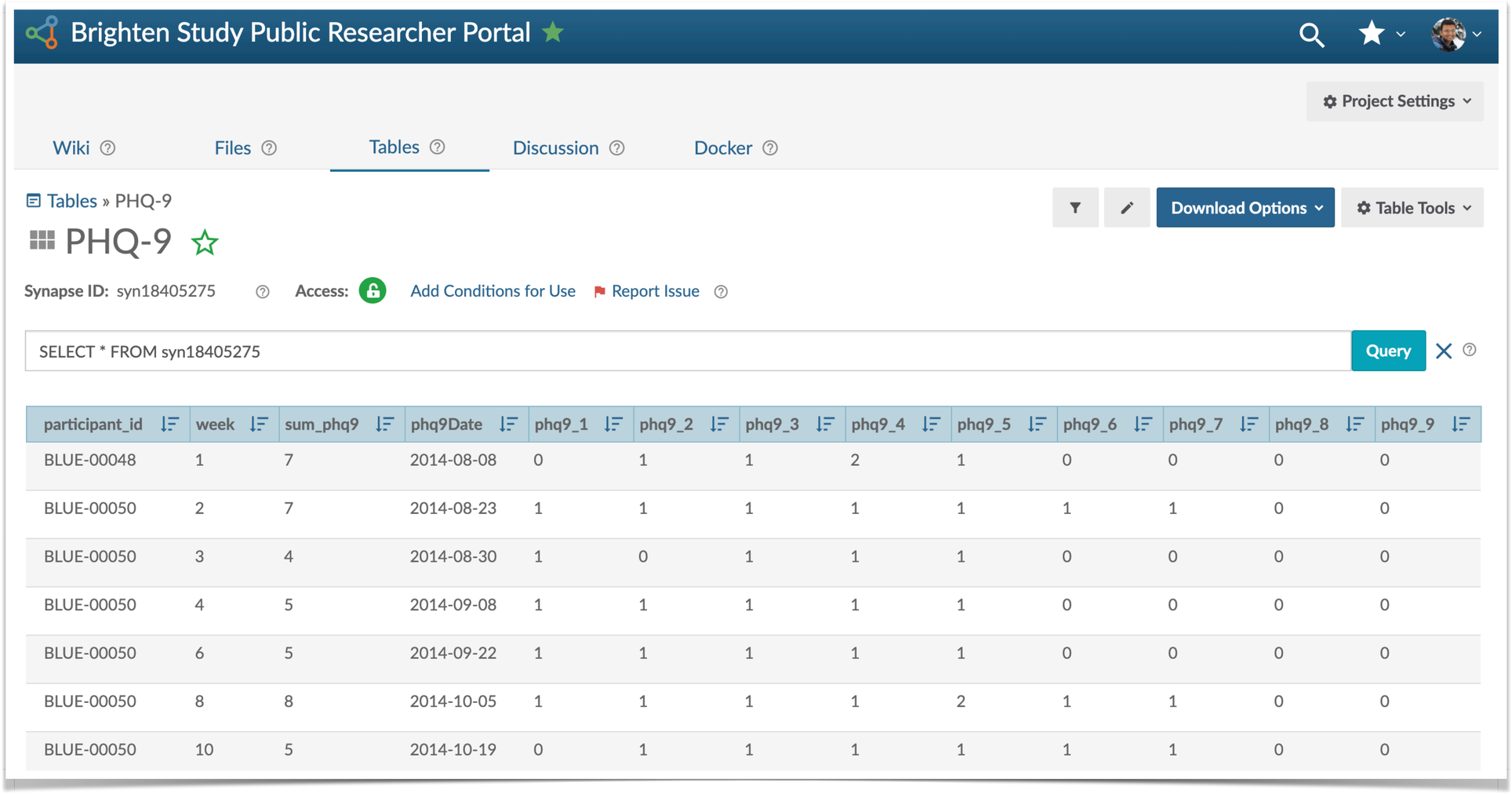
Synapse python API demo

Tutorial / Demo : bit.ly/synapse-1


clinical
omics
mobile apps
sensors
surveys
wearables
Creation
SYNAPSE
www.synapse.org
python
www
various options for Data I/O
Realtime data query & sharing
Complete data provenance
Curation
Resarch Study Portal
forum
Responsible Data Sharing
Governance
Data Annotation
Provenance Tracking
Data Queries
Data Security
Data Versioning
Consumption

Data Consumption - Example

https://www.synapse.org/brighten



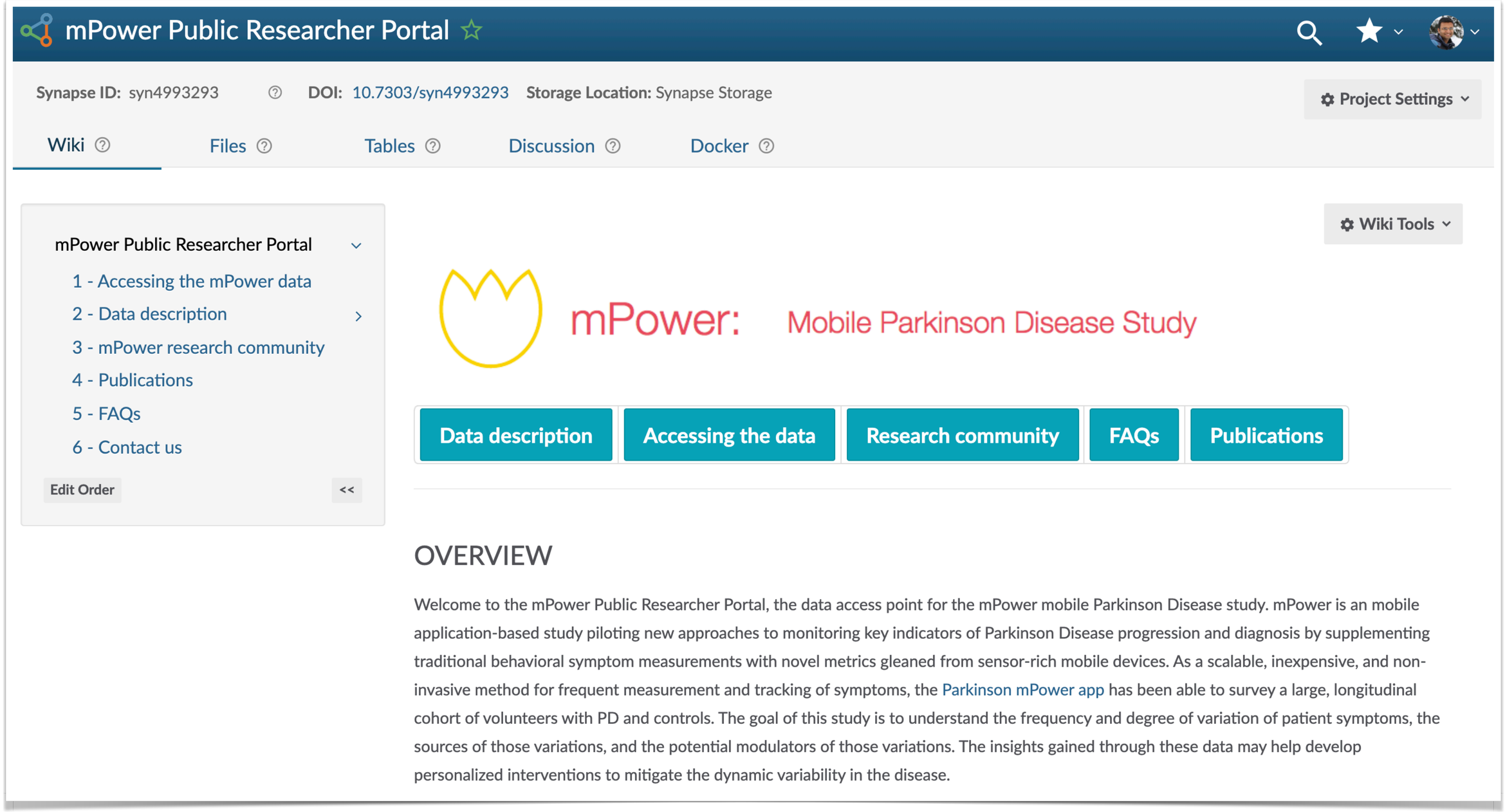
participant_id	week	sum_phq9	phq9Date	phq9_1	phq9_2	phq9_3	phq9_4	phq9_5	phq9_6	phq9_7	phq9_8	phq9_9
BLUE-00048	1	7	2014-08-08	0	1	1	2	1	0	0	0	0
BLUE-00050	2	7	2014-08-23	1	1	1	1	1	1	1	0	0
BLUE-00050	3	4	2014-08-30	1	0	1	1	1	0	0	0	0
BLUE-00050	4	5	2014-09-08	1	1	1	1	1	0	0	0	0
BLUE-00050	6	5	2014-09-22	1	1	1	1	1	0	0	0	0
BLUE-00050	8	8	2014-10-05	1	1	1	1	2	1	1	0	0
BLUE-00050	10	5	2014-10-19	0	1	1	1	1	1	1	0	0

Access Options

- Web UI
- R
- Python
- …more…

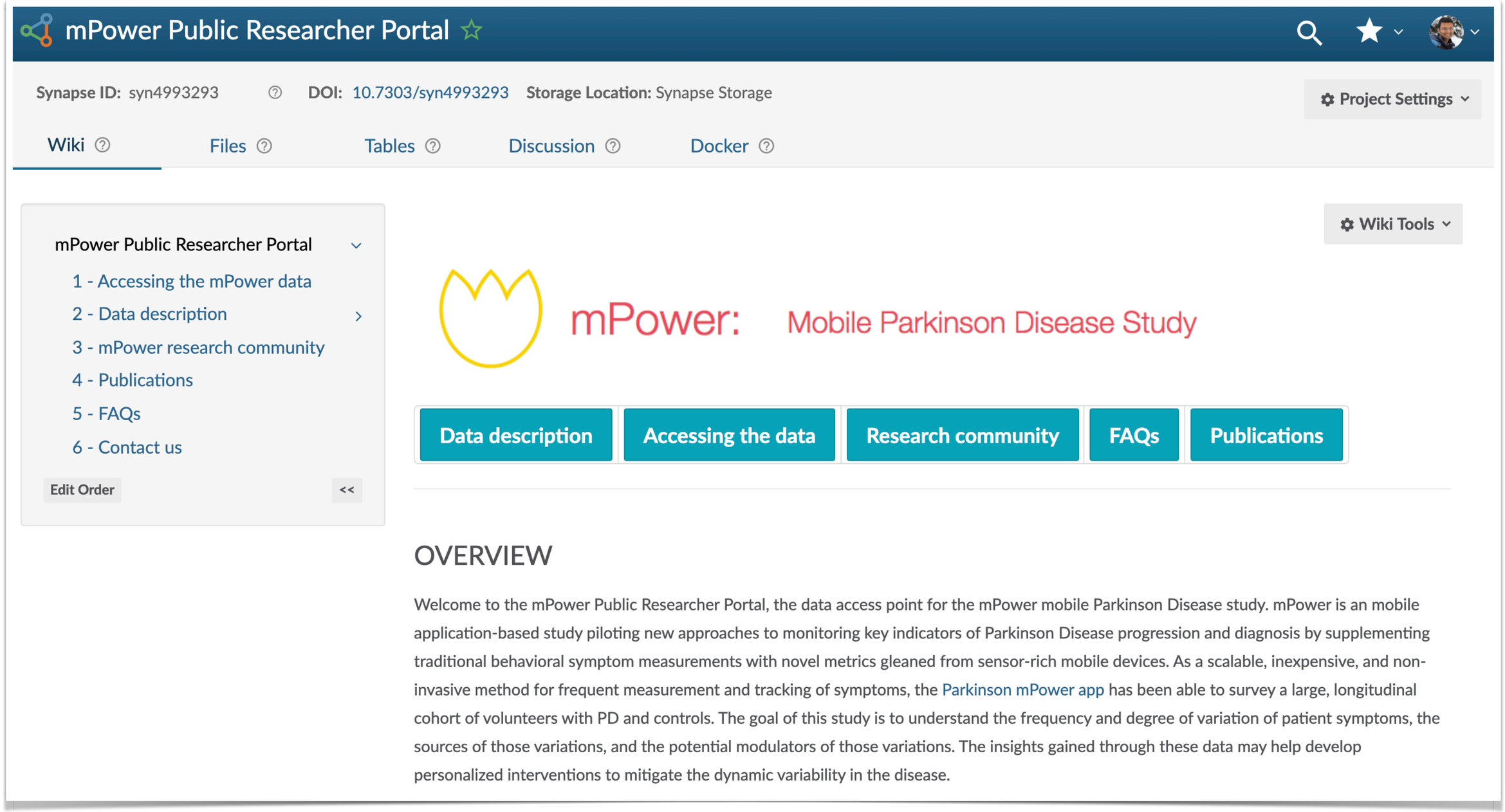
Data Consumption - Example 1

https://www.synapse.org/mpower

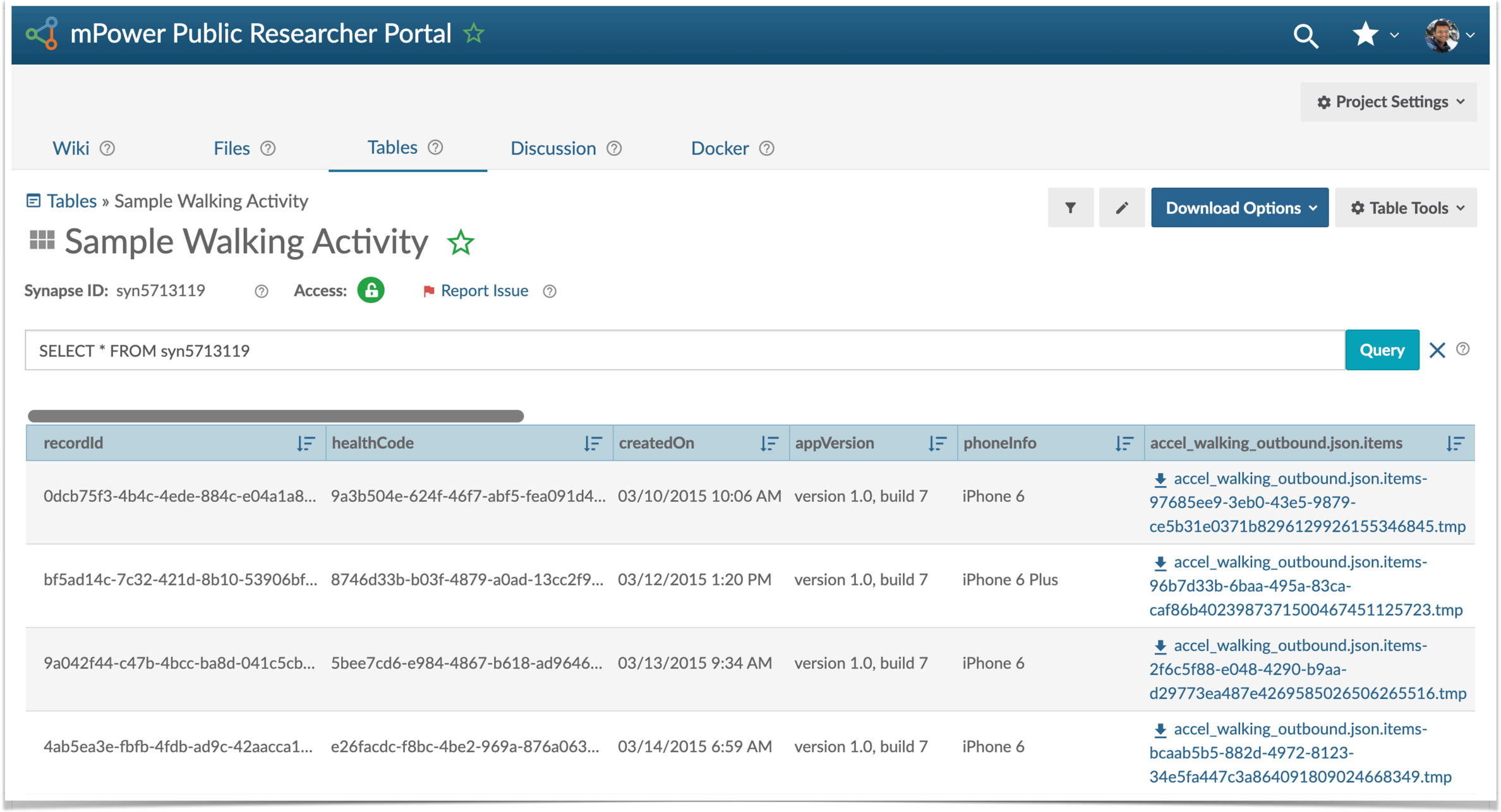


Data Consumption - Example 1

https://www.synapse.org/mpower

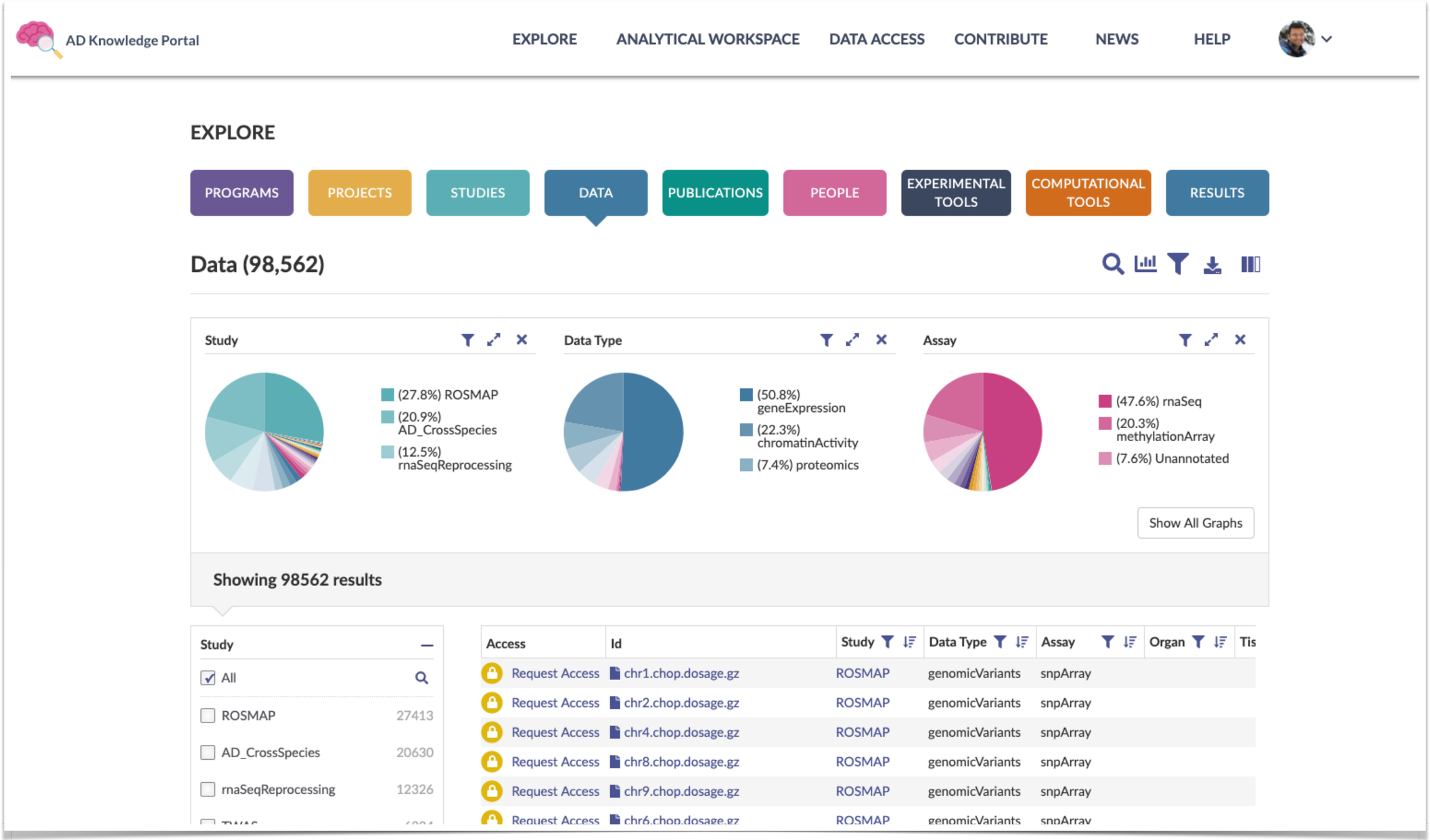


https://www.synapse.org/#!Synapse:syn5713119/tables/



Data Consumption - Example 2

https://adknowledgeportal.synapse.org/Explore/Data



dhealth.synapse.org

dHealth
Digital Health Knowledge Portal

HOME EXPLORE ABOUT

Welcome to dHealth

A Digital Health Knowledge Portal to enable the discovery and download of digital and mobile health data, tools, and benchmarked outcomes and digital biomarkers.

Supported by SAGE BIONETWORKS

Collections 10	Data 124	Tools 1	Publications 27
Studies leveraging digital health technologies, as well as analytical efforts and benchmarking challenges.	Raw and processed data from the studies catalogued in this portal.	Computational resources for the processing and analysis of digital health data.	Lessons learned and models built from digital health data, documented in peer-reviewed journal articles.
EXPLORE	EXPLORE	EXPLORE	EXPLORE

Tractable Data Analysis

mPower - Sage internal analysis working group

Alpha features on / off Trash can

Search Larsson Omberg (larssono) Help

Project Settings Tools

Wiki Files Tables Discussion Docker

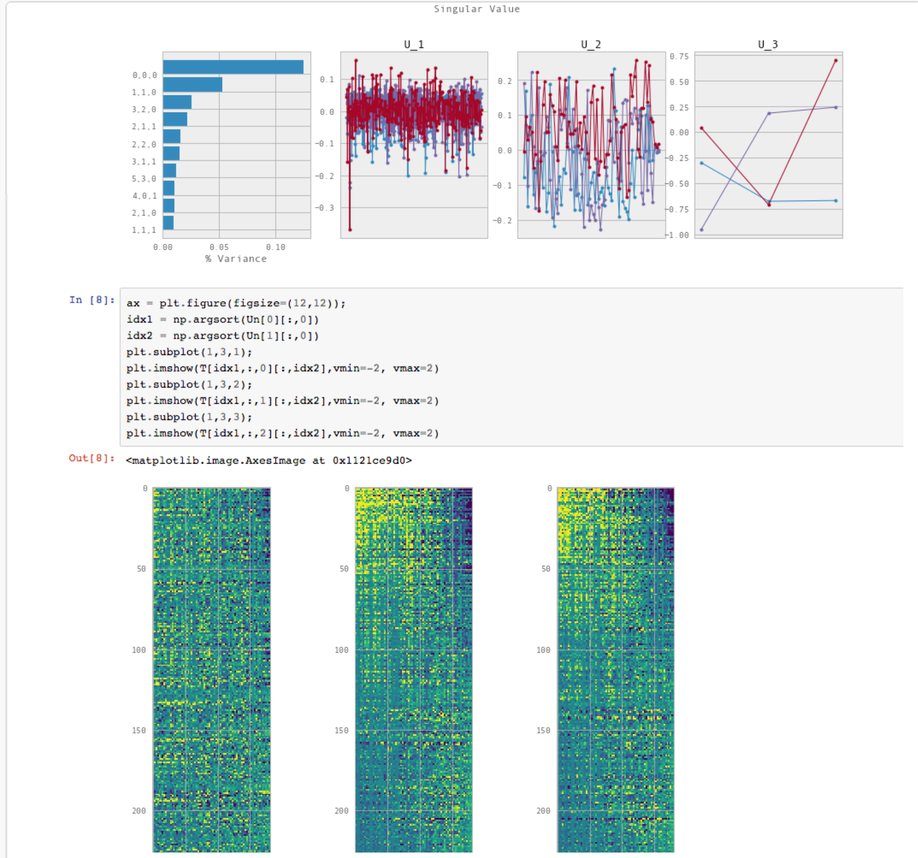
Files » analyses » Tremor » tremor_subtyping.ipynb

tremor_subtyping.ipynb Download other download options

Download file as: tremor_HOSVD.ipynb | 2.102 MB | Synapse Storage *md5*

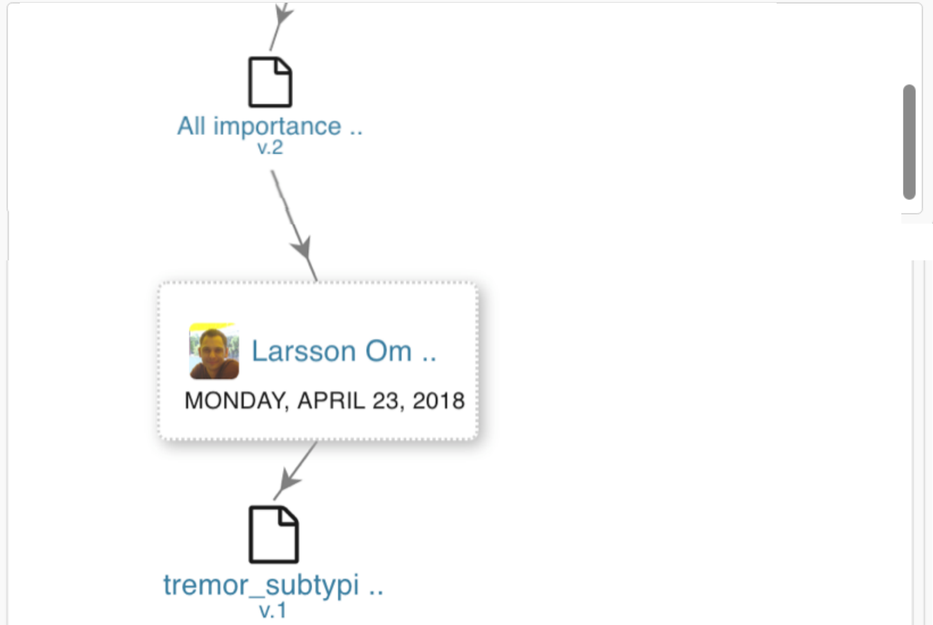
Synapse ID: syn12169237 **Conditions for use:** None (change) report issue

Preview



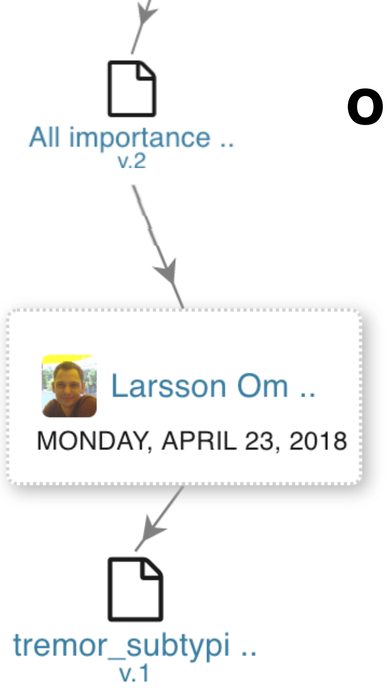

```
In [8]: ax = plt.figure(figsize=(12,12));
        idx1 = np.argsort(Un[0][:,0])
        idx2 = np.argsort(Un[1][:,0])
        plt.subplot(1,3,1);
        plt.imshow(T[idx1,:,0][:,idx2],vmin=-2, vmax=2)
        plt.subplot(1,3,2);
        plt.imshow(T[idx1,:,1][:,idx2],vmin=-2, vmax=2)
        plt.subplot(1,3,3);
        plt.imshow(T[idx1,:,2][:,idx2],vmin=-2, vmax=2)

Out[8]: <matplotlib.image.AxesImage at 0x1121ce9d0>
```


Provenance

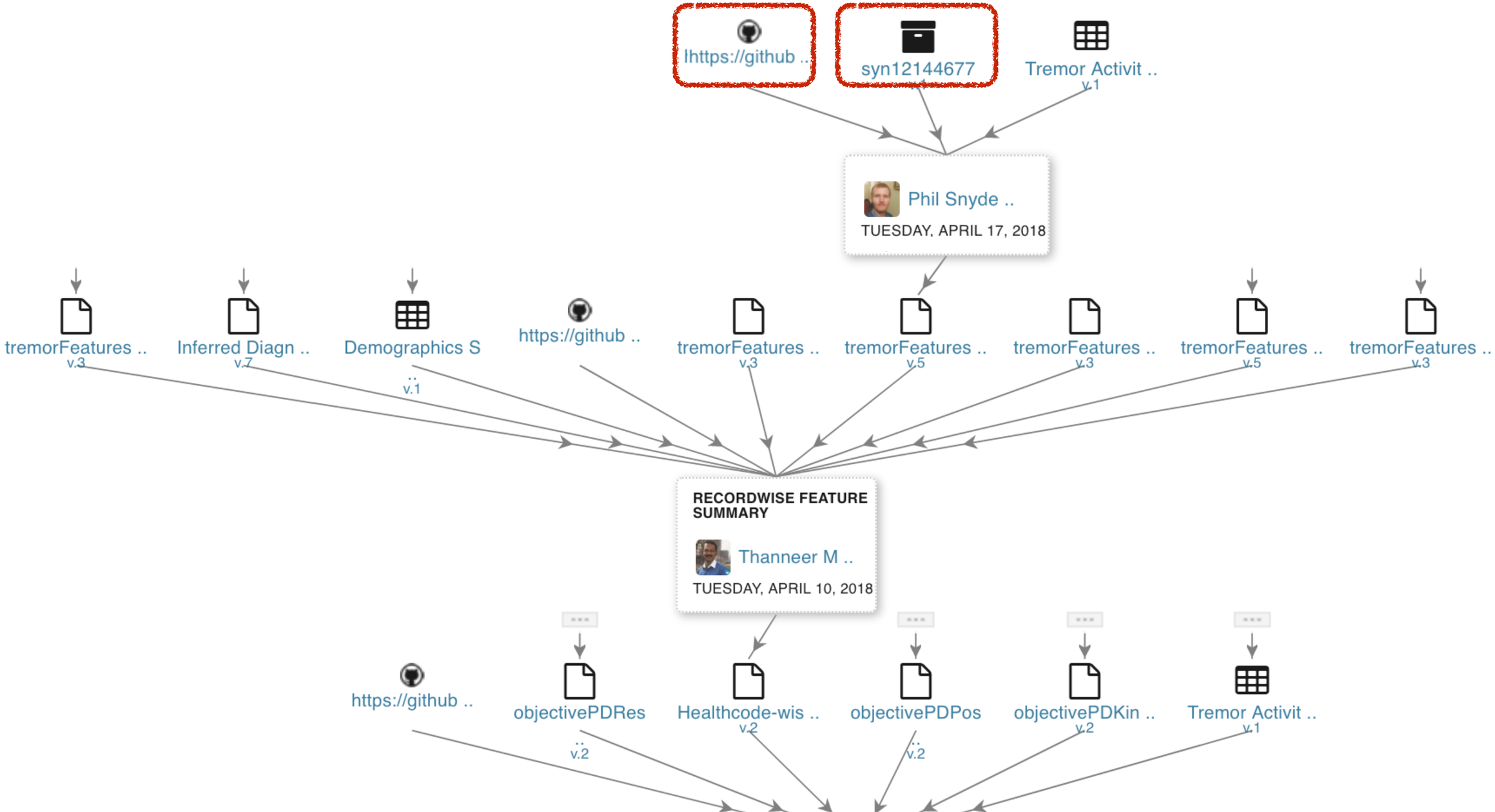


SageBionetworks

open collaborative methodologies



open collaborative methodologies



open collaborative methodologies

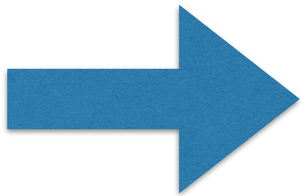
reusable

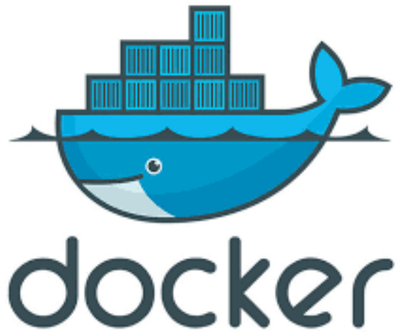
https://github

syn12144677

Sage-Bionetworks / mhealthtools

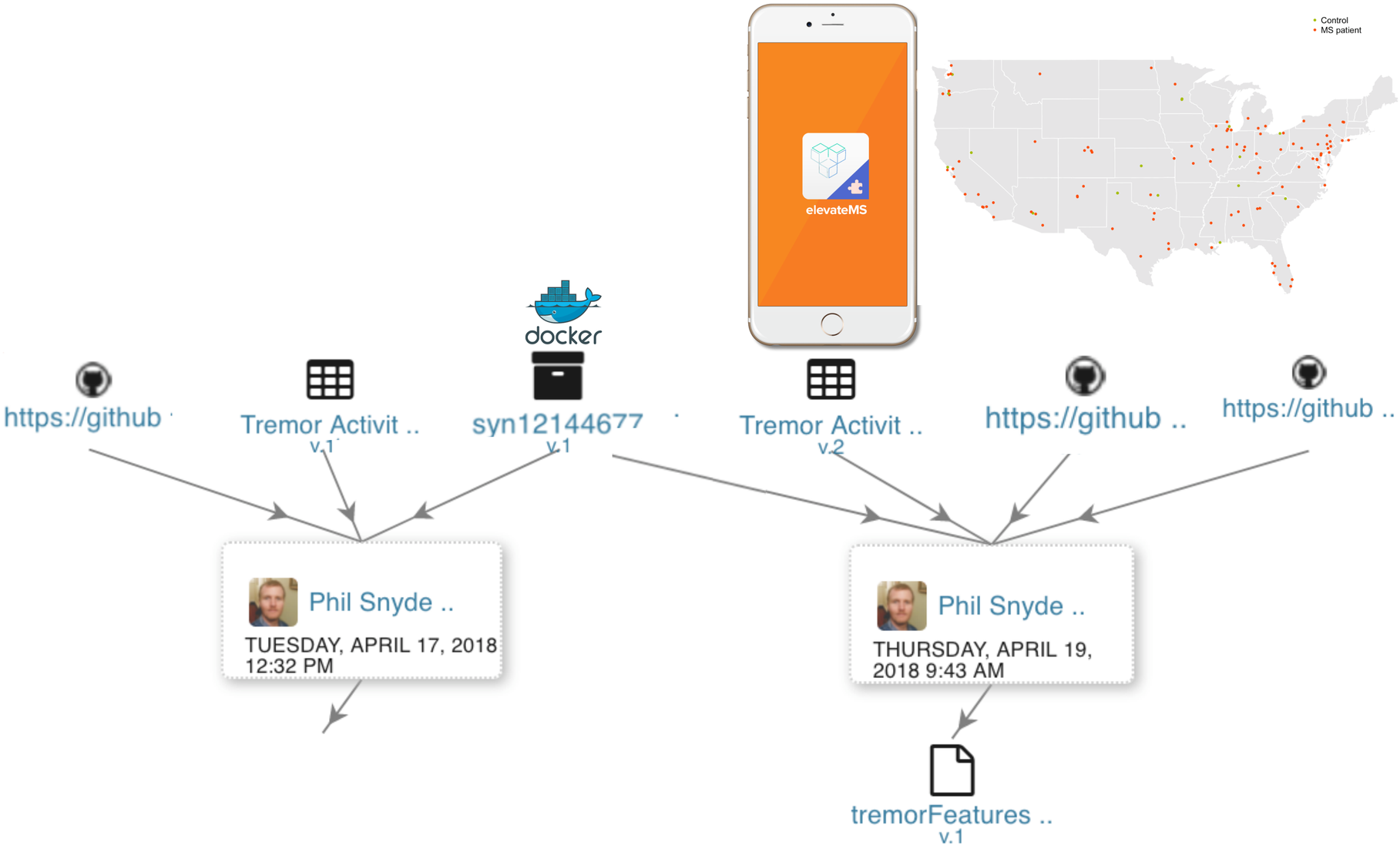
```
10  process_medicationChoiceAnswers <- functior
11      tryCatch({
12          d <- jsonlite::fromJSON(json_file)
13          data.frame(medication = paste(uniqu
14              collapse = "+"), medicationTime
15              collapse = "+"))
16      }, error = function(err) {
17          data.frame(medication = "NA", medic
18      })
```





Cloud enabled workflow

reusable



SageBionetworks

Tractable Analysis

bit.ly/synapse-2

npj | Digital Medicine

www.nature.com/npjdigitalmed

ARTICLE OPEN

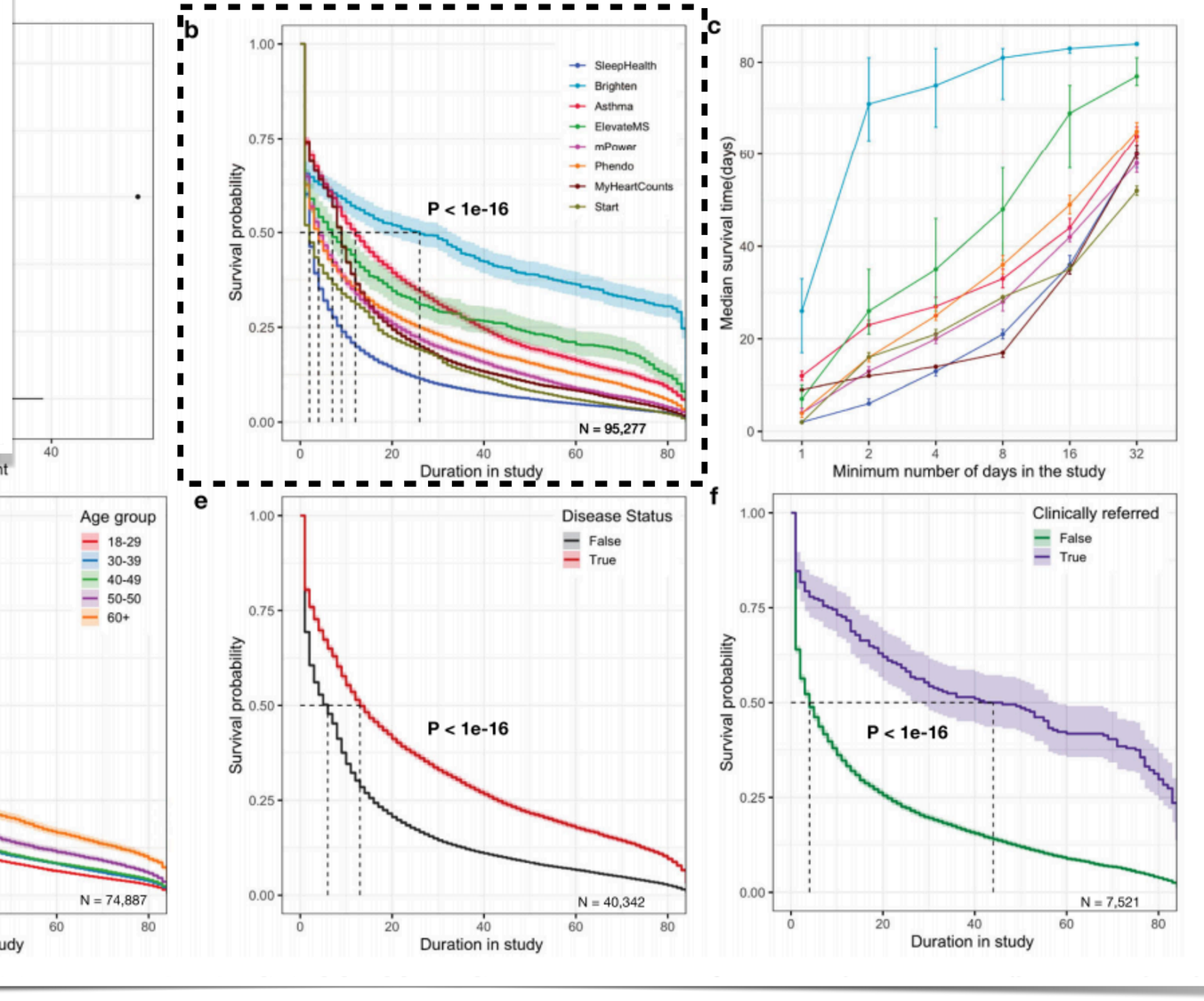
Check for updates

Indicators of retention in remote digital health studies: a cross-study evaluation of 100,000 participants

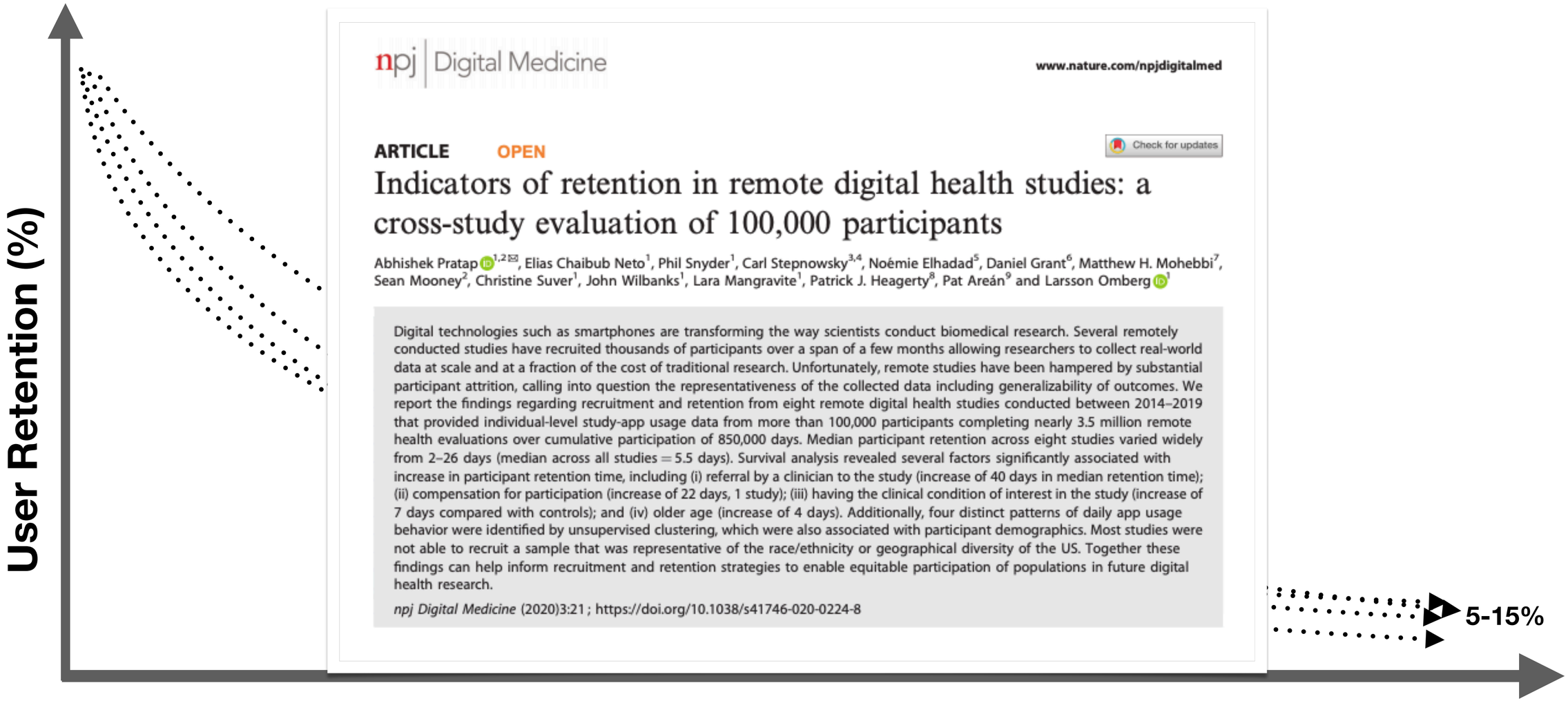
Abhishek Pratap[1,2✉], Elias Chaibub Neto[1], Phil Snyder[1], Carl Stepnowsky[3,4], Noémie Elhadad[5], Daniel Grant[6], Matthew H. Mohebbi[7], Sean Mooney[2], Christine Suver[1], John Wilbanks[1], Lara Mangravite[1], Patrick J. Heagerty[8], Pat Areán[9] and Larsson Omberg[1]

Digital technologies such as smartphones are transforming the way scientists conduct biomedical research. Several remotely conducted studies have recruited thousands of participants over a span of a few months allowing researchers to collect real-world data at scale and at a fraction of the cost of traditional research. Unfortunately, remote studies have been hampered by substantial participant attrition, calling into question the representativeness of the collected data including generalizability of outcomes. We report the findings regarding recruitment and retention from eight remote digital health studies conducted between 2014–2019 that provided individual-level study-app usage data from more than 100,000 participants completing nearly 3.5 million remote health evaluations over cumulative participation of 850,000 days. Median participant retention across eight studies varied widely from 2–26 days (median across all studies = 5.5 days). Survival analysis revealed several factors significantly associated with increase in participant retention time, including (i) referral by a clinician to the study (increase of 40 days in median retention time); (ii) compensation for participation (increase of 22 days, 1 study); (iii) having the clinical condition of interest in the study (increase of 7 days compared with controls); and (iv) older age (increase of 4 days). Additionally, four distinct patterns of daily app usage behavior were identified by unsupervised clustering, which were also associated with participant demographics. Most studies were not able to recruit a sample that was representative of the race/ethnicity or geographical diversity of the US. Together these findings can help inform recruitment and retention strategies to enable equitable participation of populations in future digital health research.

npj Digital Medicine (2020)3:21 ; https://doi.org/10.1038/s41746-020-0224-8



Building digital tech alone is not enough



npj | Digital Medicine

www.nature.com/npjdigitalmed

ARTICLE OPEN

Check for updates

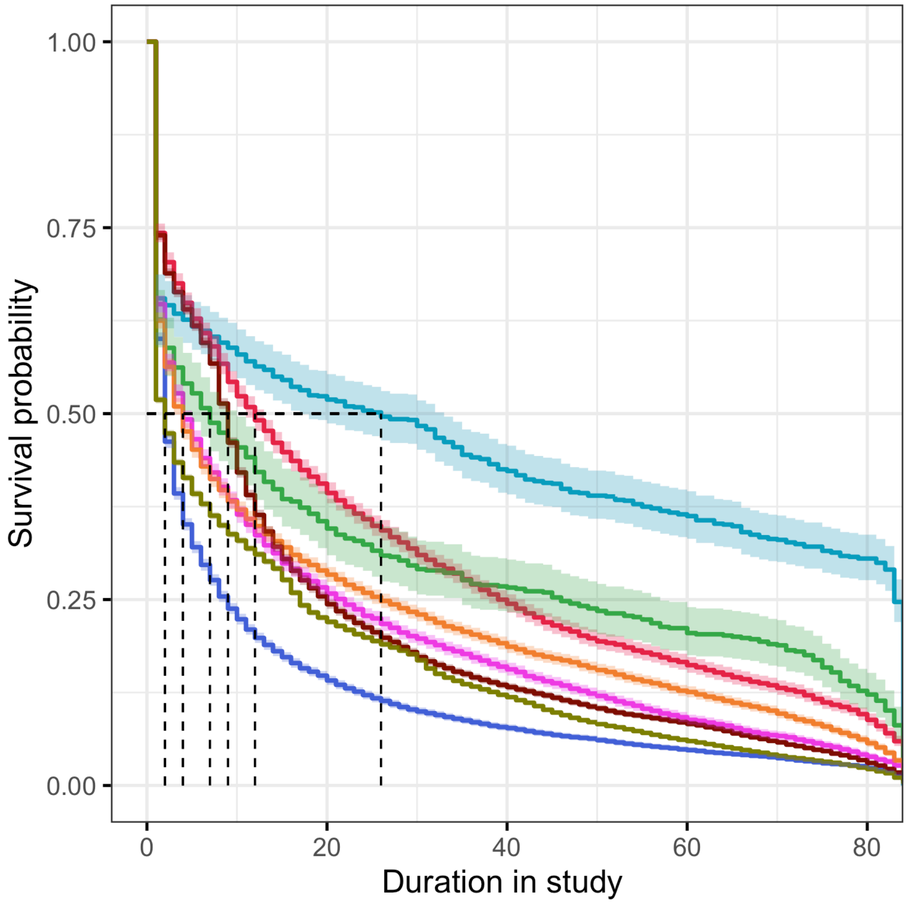
Indicators of retention in remote digital health studies: a cross-study evaluation of 100,000 participants

Abhishek Pratap[1,2✉], Elias Chaibub Neto[1], Phil Snyder[1], Carl Stepnowsky[3,4], Noémie Elhadad[5], Daniel Grant[6], Matthew H. Mohebbi[7], Sean Mooney[2], Christine Suver[1], John Wilbanks[1], Lara Mangravite[1], Patrick J. Heagerty[8], Pat Areán[9] and Larsson Omberg[1]

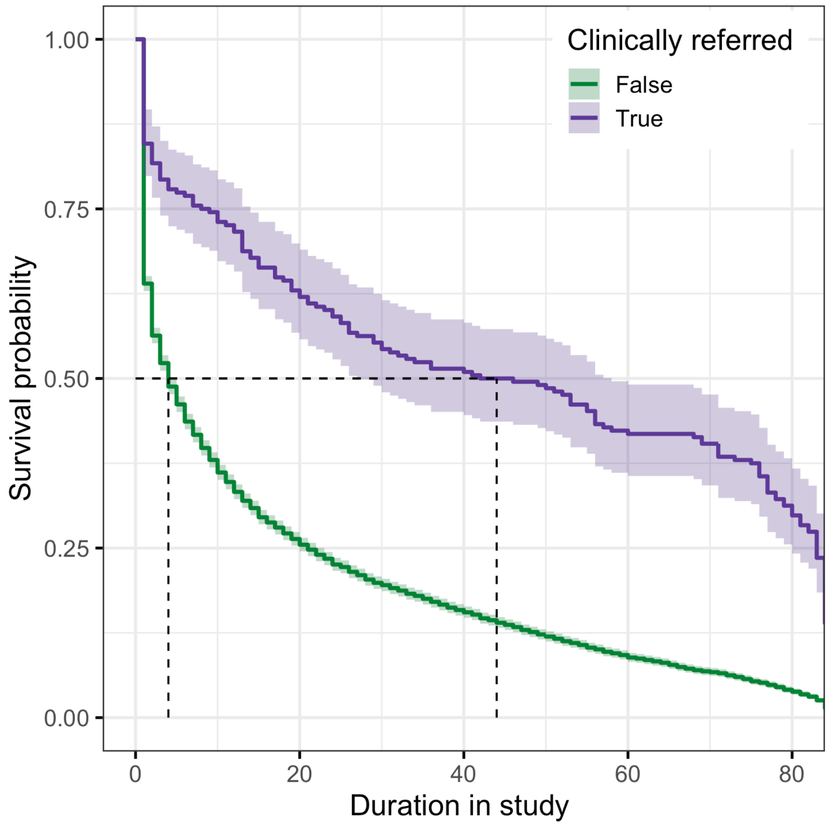
Digital technologies such as smartphones are transforming the way scientists conduct biomedical research. Several remotely conducted studies have recruited thousands of participants over a span of a few months allowing researchers to collect real-world data at scale and at a fraction of the cost of traditional research. Unfortunately, remote studies have been hampered by substantial participant attrition, calling into question the representativeness of the collected data including generalizability of outcomes. We report the findings regarding recruitment and retention from eight remote digital health studies conducted between 2014–2019 that provided individual-level study-app usage data from more than 100,000 participants completing nearly 3.5 million remote health evaluations over cumulative participation of 850,000 days. Median participant retention across eight studies varied widely from 2–26 days (median across all studies = 5.5 days). Survival analysis revealed several factors significantly associated with increase in participant retention time, including (i) referral by a clinician to the study (increase of 40 days in median retention time); (ii) compensation for participation (increase of 22 days, 1 study); (iii) having the clinical condition of interest in the study (increase of 7 days compared with controls); and (iv) older age (increase of 4 days). Additionally, four distinct patterns of daily app usage behavior were identified by unsupervised clustering, which were also associated with participant demographics. Most studies were not able to recruit a sample that was representative of the race/ethnicity or geographical diversity of the US. Together these findings can help inform recruitment and retention strategies to enable equitable participation of populations in future digital health research.

npj Digital Medicine (2020)3:21 ; https://doi.org/10.1038/s41746-020-0224-8

A Pratap, E Neto, P Snyder, C Stepnowsky, N Elhadad, D Grant, M Mohebbi, S Mooney, C Suver, J Wilbanks, L Mangravite, P Heagerty, P Arean, L Omberg - Indicators of retention in remote digital health studies: A cross-study evaluation of 100,000 participants **| Nature Digital Medicine, 2020**



- ~50% of participants leave the studies within the first 7-10 days
- Targeted Comms during the first week may help



- Participants referred by clinical sites/partners engage for significantly longer time