ABCD-ReproNim: An ABCD Course on Reproducible Data Analyses

ABCD: Data Exploration and Analysis Portal (ABCD DEAP)

Wes Thompson

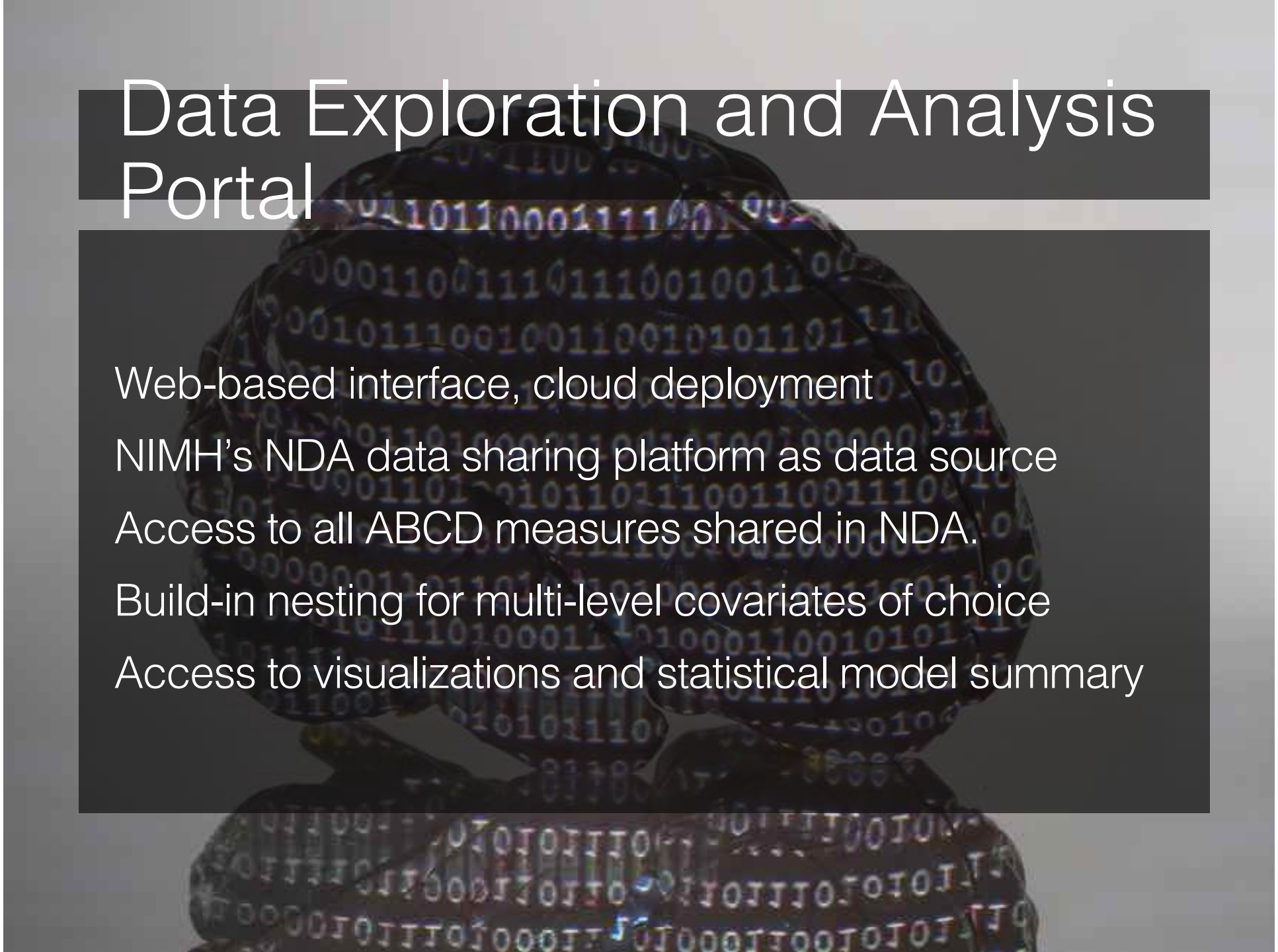
Department of Family Medicine and Public Health
Division of Biostatistics
University of California, San Diego

wkthompson@health.ucsd.edu

Learning Objectives of this Lecture

- Learn about the ABCD study design
- Learn basic analysis that incorporate the ABCD Study design
- Learn how to access the data and perform analyses in DEAP

Data Exploration and Analysis Portal



Web-based interface, cloud deployment

NIMH's NDA data sharing platform as data source

Access to all ABCD measures shared in NDA.

Build-in nesting for multi-level covariates of choice

Access to visualizations and statistical model summary

ABCD open science

ABCD-ReproNim

[https://github.com/ABCD-STUDY/]

enroll
Participant enrollment system hosting sensitive data
PHP Updated 7 hours ago

DEAP
Data Exploration and Analysis Portal of the ABCD Study
JavaScript ★ 2 1 Updated a day ago

analysis-nda17
Collection of scripts to analyze ABCD release data
nda17 abcd-study
★ 6 5 Updated 2 days ago

redcap_rewrite_history
Change the name of REDCap items in an existing project and attempt to rewrite the projects history.
PHP ★ 1 GPL-3.0 Updated 5 days ago

FIONASITE
Data upload site for FIONA site computer
JavaScript 1 Updated 7 days ago

complete_row
REDCap extension module: Colors each row of an instrument - if a value has been provided. This will highlight rows with missing values to improve their visibility.
PHP GPL-3.0 Updated on Aug 20

CIFTI-Analysis
Scripts to enable vertex-wise (CIFTI) analysis of ABCD with FSL/PALM and HCP Workbench
Python ★ 1 Updated on Aug 16

redcap-to-nda
Exporting REDCap data dictionaries and data to the NIMH National Data Archive (NDA)
JavaScript 1 Updated on Jul 31

auto-scoring
Visual programming to calculate derived scores for REDCap
JavaScript Updated on Jun 29

eprime-data-clean
Convert E-Prime generated files (exported as csv) to propper CSV
Python Updated on Jun 25

ABCDWorkshop-reproducible-science
Material collection for the ABCD-DAIC workshops on data science
JavaScript Updated on Jun 16

timeline-followback
Online timeline-followback subject test
JavaScript Updated on Apr 24

nih-ipad-app-end-point
An end-point for centrally storing data from the NIH iPad app
PHP 1 Updated on Apr 20

geocoding
A framework for adding gelocation derived data to the ABCD study.
R MIT Updated on Apr 4

FIONA-QC-PHANTOM

Online QC operations performed on Phantom MRI data

Matlab ★1 1 Updated on Mar 23

Fast-Track-Image-Sharing

The ABCD study shares data on the National Data Archive. This project provides the tools for sharing.

nda dicom-images anonymization

Python ★1 Updated on Feb 7

Minimally-Processed-Image-Sharing

Python ★2 Updated on Dec 5, 2017

little-man-task

The little man task web-based instrument

JavaScript Updated on Nov 30, 2017

redcap-completion

Measure item level completion in a large REDCap project

JavaScript Updated on Nov 11, 2017

simple-t1-motion-detection

Measures the amount of ghosting artifacts in T1-weighted images

C++ Updated on Jul 27, 2017

tick-tock

Study Observation system monitoring events per day

JavaScript Updated on Jun 23, 2017

numerical-fitting

Client side numerical computation library written in javascript.

JavaScript Updated on Dec 28, 2016

aux-file-upload

PHP Updated on Dec 2, 2016

FIONA-protocol-compliance

Matlab script for ABCD study protocol compliance

Matlab Updated on Nov 28, 2016

redcap-hook-framework

Forked from 123andy/redcap-hook-framework

The REDCap hook framework is a means to organize and deploy custom hooks in a single project or across the entire instance.

PHP ★1 16 Updated on Nov 4, 2016

ABCDreport

PHP Updated on Sep 6, 2016

pearson-central-end-point

An end-point for centrally storing data from the Pearson's Q-interactive.

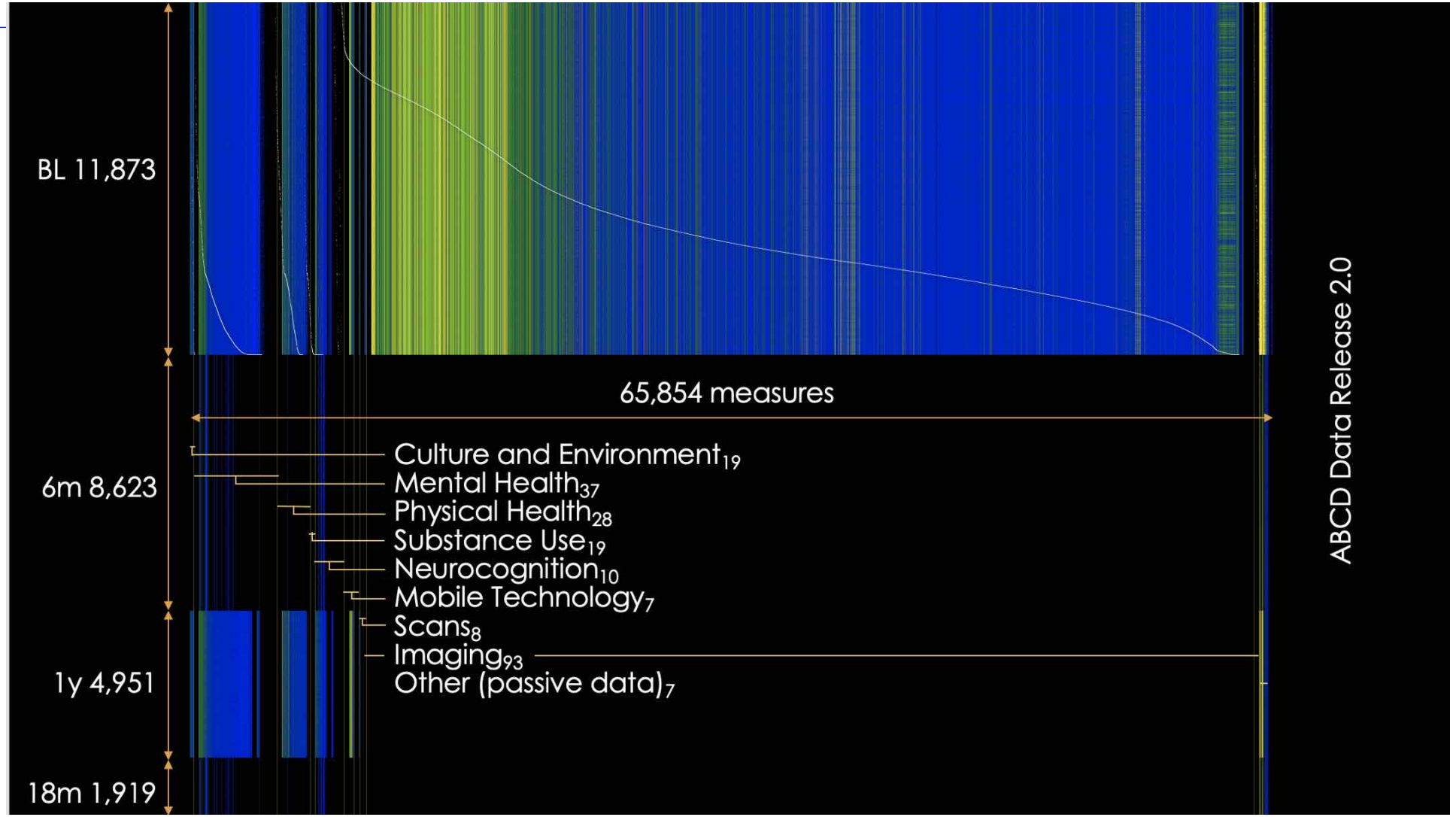
PHP Updated on Jun 7, 2016

delay-discounting

Delay-discounting task measuring impulsivity

JavaScript Updated on Aug 11, 2016

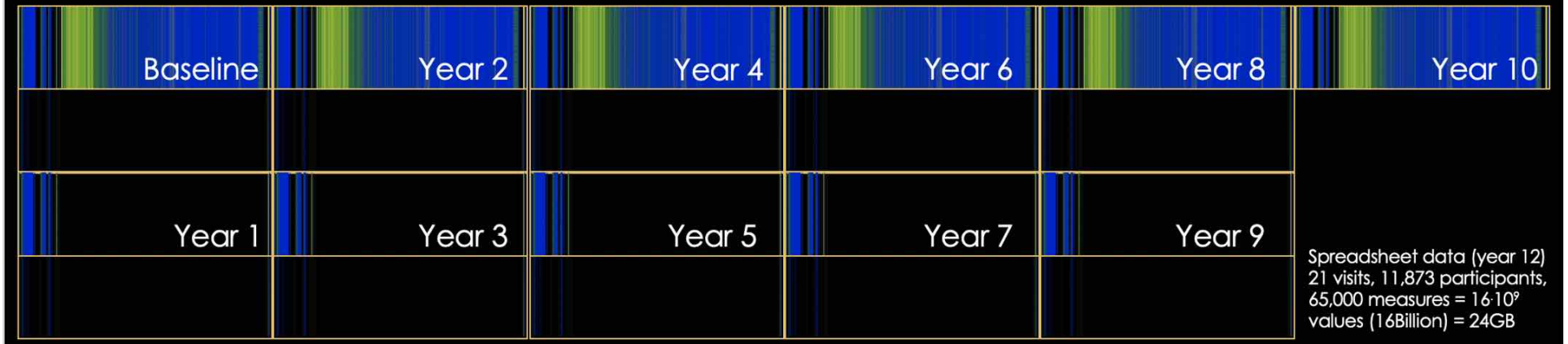
ABCD STUDY DESIGN



ABCD STUDY DESIGN

Release Year	Baseline	6 month	1 year	18 month	2 year	36 month	3 year	48 month	4 year	60 month	5 year	72 month	6 year	84 month	7 year	96 month	8 year	108 month	9 year	120 month	10 year	132 month	11 year	144 month	12 year
1	4,951	0																							
2	11,873	8,623	4,951	1,919	0																				
3	11,873	11,873	11,873	8,905	5,937	2,968	0																		
4	11,873	11,873	11,873	11,873	11,873	8,905	5,937	2,968	0																
5	11,873	11,873	11,873	11,873	11,873	11,873	11,873	8,905	5,937	2,968	0														
6	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	8,905	5,937	2,968	0												
7	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	8,905	5,937	2,968	0										
8	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	8,905	5,937	2,968	0								
9	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	8,905	5,937	2,968	0						
10	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	8,905	5,937	2,968	0				
11	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	8,905	5,937	2,968	0		
12	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	8,905	5,937	2,968	0
13	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	8,905	5,937
14	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873	11,873
Collection year	0	0.5	1	1.5	2	2.5	3	3.5	4	4.5	5	5.5	6	6.5	7	7.5	8	8.5	9	9.5	10	10.5	11	11.5	12

Yearly (rolling) release schedule



Spreadsheet data (year 12)
21 visits, 11,873 participants,
65,000 measures = $16 \cdot 10^9$
values (16Billion) = 24GB

https://www.biorxiv.org/content/10.1101/2020.09.01.276451v1

New Results

Comment on this paper

Meaningful Effects in the Adolescent Brain Cognitive Development Study

Anthony Steven Dick, Ashley L. Watts, Steven Heeringa, Daniel A. Lopez, Hauke Bartsch, Chun Chieh Fan, Clare Palmer, Chase Reuter, Andrew Marshall, Frank Haist, Samuel Hawes, Thomas E. Nichols, Deanna M. Barch, Terry L. Jernigan, Hugh Garavan, Steven Grant, Vani Pariyadath, Elizabeth Hoffman, Michael Neale, Martin P. Paulus, Kenneth J. Sher, Wesley K. Thompson

doi: https://doi.org/10.1101/2020.09.01.276451

Abstract

The Adolescent Brain Cognitive Development (ABCD) Study is the largest single-cohort prospective longitudinal study of neurodevelopment and children's health in the United States. A cohort of n= 11,880 children aged 9-10 years (and their parents/guardians) were recruited across 22 sites and are being followed with in-person visits on an annual basis for at least 10 years. The study approximates the US population on several key sociodemographic variables, including sex, race, ethnicity, household income, and parental education. Data collected include assessments of health, mental health, substance use, culture and environment and neurocognition, as well as geocoded exposures, structural and functional magnetic resonance imaging (MRI), and whole-genome genotyping. Here, we describe the ABCD Study aims and

ABCD Adolescent Brain Cognitive Development
Data Exploration and Analysis Portal

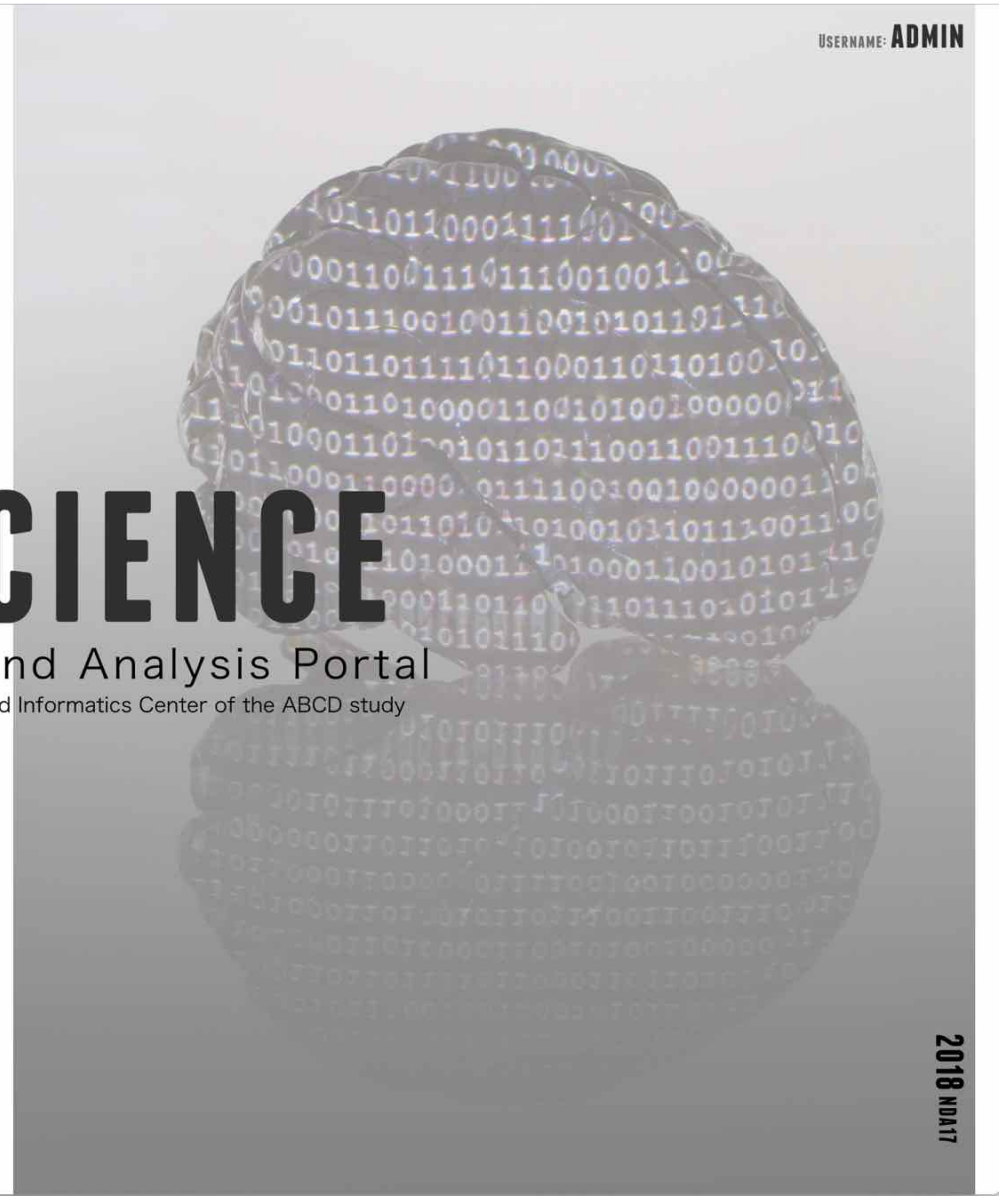
USERNAME: ADMIN

GETTING STARTED
00 PLAN
01 EXPLORE
02 LIMIT
03 ANALYSE
04 EXTEND

DEAP SCIENCE

Data Exploration and Analysis Portal

A service provided by the Data Analysis and Informatics Center of the ABCD study

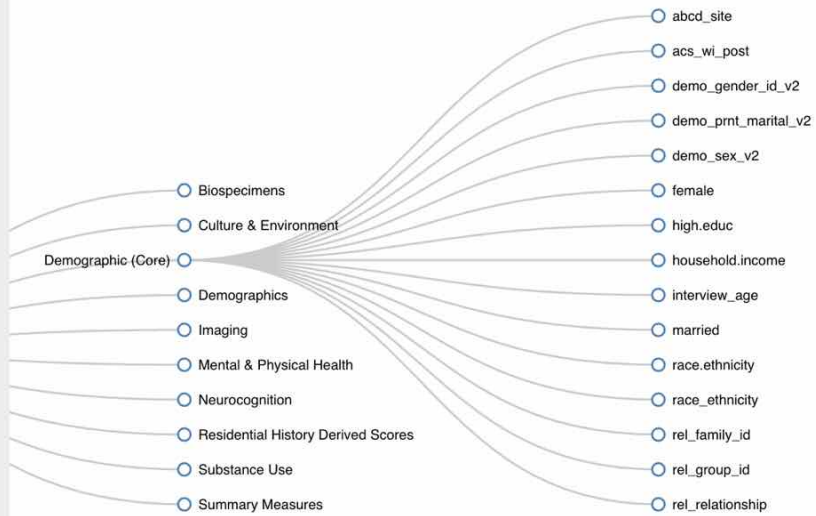


2018 NDA17

ABCD Ontology Viewer Home

ABCD Ontology

click to expand or collapse, drag to pan, scroll-wheel to zoom



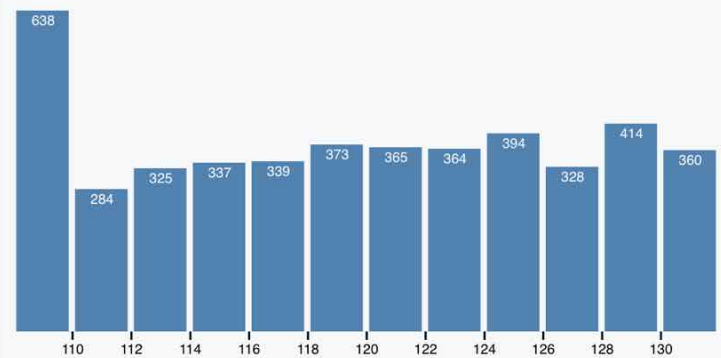
interview_age ?

examples: intelligence, schizophrenia, ADHD

More than 101 results (0.11 seconds)

interview_age in ABCD Children's Report of Parental Behavioral Inventory / crpbi01 [Parenting]

```
Min.  108
1st Qu.  114
Median  120
Mean  120.02
3rd Qu.  126
Max.  131
```



[Open NDA for crpbi01 (new tab)]

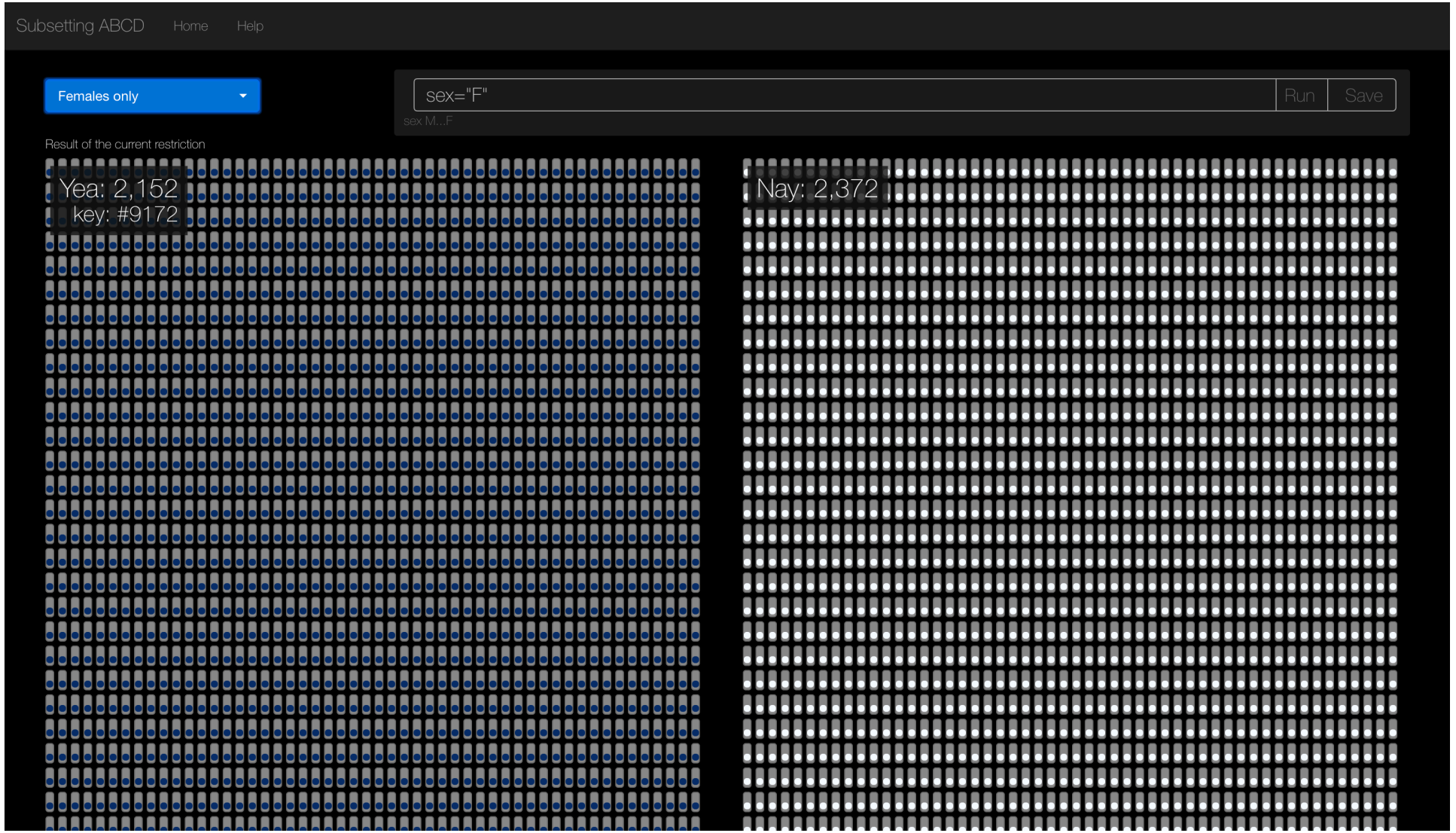
search term: interview_age - matches element name

Age in months at the time of the interview/test/sampling/imaging.

Age is rounded to chronological month. If the research participant is 15-days-old at time of interview, the appropriate value would be 0 months. If the participant is 16-days-old, the value would be 1 month.

interview_age in ABCD Cash Choice Task / cct01 [Task Based]

search term: interview_age - matches element name


Subsetting ABCD
Home
Help
Females only
sex="F"
Run
Save
sex M...F
Result of the current restriction
Yea: 2,152
key: #9172
Nay: 2,372

[admin] bmi_calc_example The body mass index calculated from the height and weight va… close

Element Name (user admin - public score)

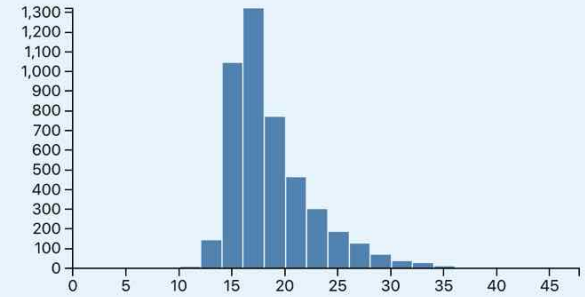
bmi_calc_example

Axis label

The body mass index calculated from the height and weigh

Save as private Save

bmi_calc_example	anthroheightcalc	anthroweightcalc	eventname
18.234	56.5	82.8	baseline_year
20.15	56.5	91.5	baseline_year
15.174	57.3	70.8666666666667	baseline_year
19.993	53.5	81.4	baseline_year
17.663	58.3	85.4	baseline_year
16.213	54.5	68.5	baseline_year
20.468	55.35	89.2	baseline_year
34.171	63.5	196	baseline_year



The Body-Mass Index (BMI)

The body-mass-index (BMI) depends on the height and weight of the participant. These two values exist for each participant in DEAP. We need to copy these values by calling the `use`-function into our browser window. As a return value `use` returns a list of promises that are fulfilled once all the data arrives.

```
var promises = use(["anthroweightcalc", "anthroheightcalc"]);
```

The BMI can be calculated using the following formula - assuming pounds as units for weight (w) and inches as units for height (h):

$$703\frac{w}{h^2}$$

We can implement this calculation in a function called `calc` that gets two arguments, the weight of a participant in `w` and the height of a participant in `h`. The function then returns the calculated values.

```
function calc(w, h) {
     return w/(h*h) * 703;
}
```

Now we wait until the promises have been resolved, which indicates that the weight and height values are available. At this point we can get the data and compute the new variable `anthro_weight_calc` using map. The map function computes for each row of the data spreadsheet the value of the new variable. It is sufficient to row.set the new value to have it show up in the histogram and table of this variable:

```
Promise.all(promises).then(function() {
    var data = new DataFrame(allMeasures);
```

Autosaved: 2:30 pm lines: 27 words: 297 0:0 2 Keystrokes Cmd-P

Multilevel Data Analysis

Multilevel statistical models for baseline data reflect the multilevel study design (GAMM4).

$$Y_{sfi} = \beta_0 + \boldsymbol{x}_{sfi}\boldsymbol{\beta} + \boldsymbol{z}_{sfi}\boldsymbol{\gamma} + a_s + b_{f_{(s)}} + \epsilon_{sfi}$$

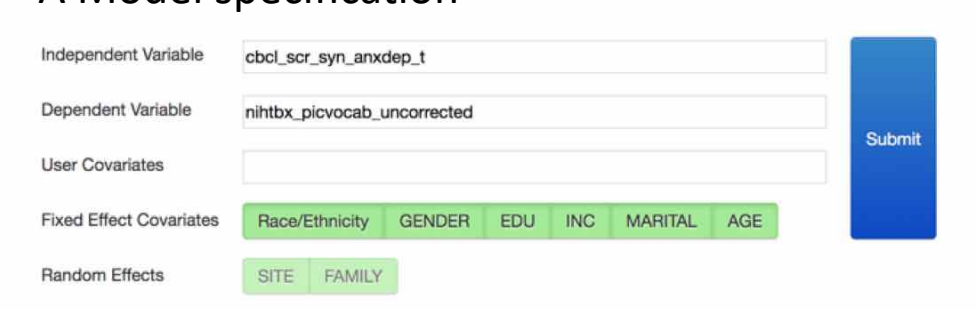
- x_{sfi} are covariates (e.g., demographics)
- z_{sfi} are independent variables of interest
- a_s is a site-specific random effect
- $b_{f(s)}$ is a family random effect nested within site

This model is extendable to non-normal outcomes (e.g., discrete, count variables).

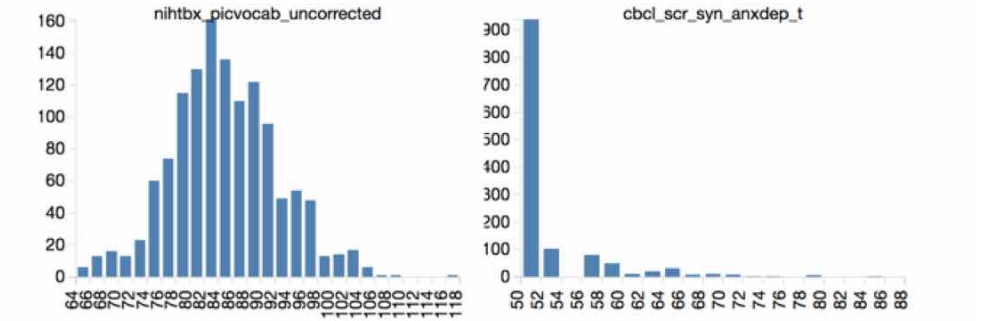
Model Fitting with DEAP

Can changes in anxiety be explained by cognitive development scores measured in the picture vocabulary test, if one corrects for known covariates?

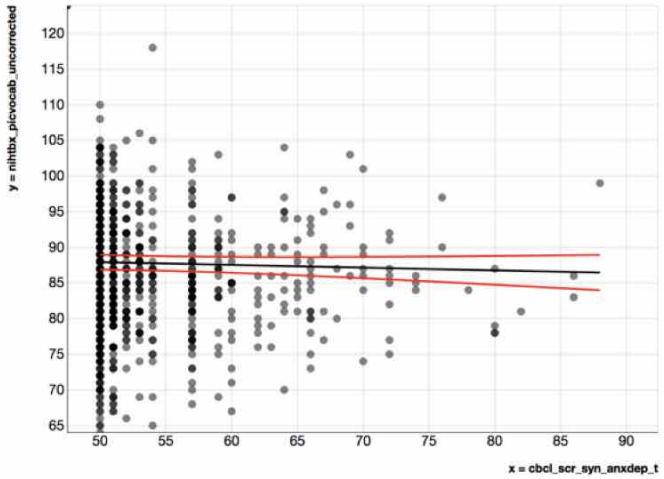
A Model specification



B Data used in the model



C Regression model fit



D Result tables / Model comparisons

	Estimate	Std. Error	t value	Pr(>|t|)	sig
(Intercept)	52.27064	1.77974	29.37	< 1e-6	***
nihtbx_picvocab_uncorrected	0.02316	0.01322	1.75	0.0798201	.
race.ethnicityBlack	-1.15741	0.37474	-3.09	0.0020246	**
race.ethnicityHispanic	-0.14640	0.30244	-0.48	0.628372	
race.ethnicityAsian	-1.21511	0.66369	-1.83	0.0671952	.
race.ethnicityOther	0.13576	0.33444	0.41	0.6848096	
genderM	0.67781	0.18458	3.67	0.0002436	***
high.educBachelor	-0.05391	0.54923	-0.10	0.9218111	
high.educHS Diploma/GED	-0.90738	0.57636	-1.57	0.1154924	
high.educPost Graduate Degree	-0.17039	0.56453	-0.30	0.7628061	
high.educSome College	-0.06243	0.52201	-0.12	0.9048016	
marriedyes	-0.40629	0.24155	-1.68	0.0926505	.
interview_age	-0.00946	0.01301	-0.73	0.4672105	
household.income[< 50K]	1.12847	0.32764	3.44	0.0005784	***
household.income[> =50K& < 100K]	0.48843	0.24194	2.02	0.0435734	*

Table 3: Statistical parameter table.

Tutorial Mode on DEAP

Not familiar with generalized additive mixed models for the analysis of longitudinal data in a multi-site project with a complex family structure? Deap provides a training-wheel mode with in-depth explanations on how to interpret your model.

an independent variable again. Use the buttons to toggle off the inclusion of any of the fixed effect covariates. Both site and family are always included into the model as random effects as they are part of the study design.

Data Display and Summaries

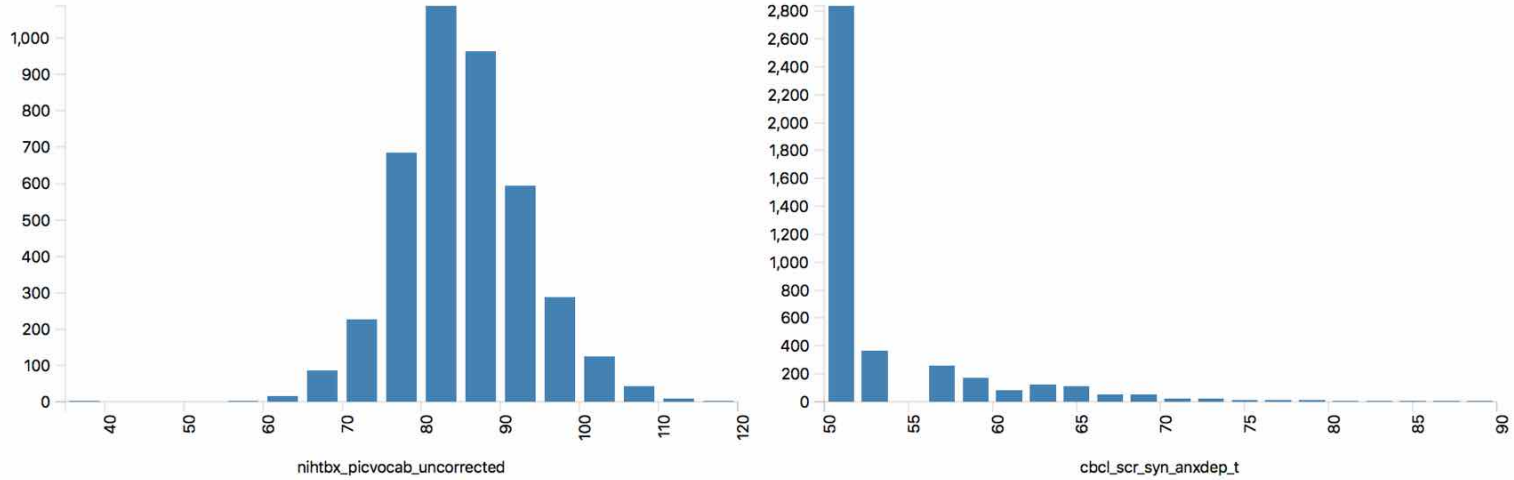


Fig. 2: Data distributions for dependent (left) and independent variable (right).

Histograms are used to inspect the distributions of the data used in the model. For the dependent variable (Fig. 2, left) we want to make sure that they are roughly normally distributed (bell-shaped). In particular we want to check if there are outliers or, if the distribution is highly skewed. If large

Feature: Expert Mode

Access to the (R) source code behind the GAMM4 model. Can be edited by the user and becomes part of a sharable resource for download and to other DEAP users.

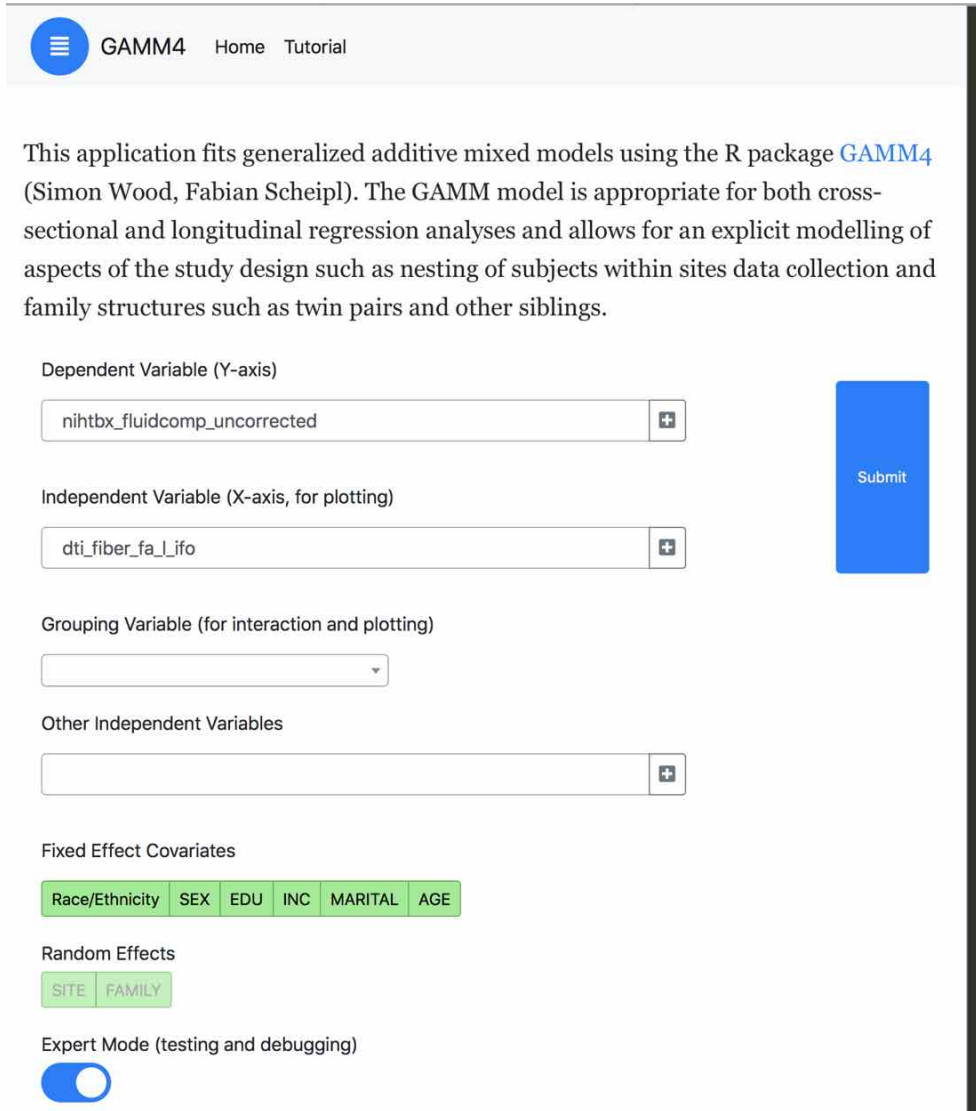


Fig. 1: Model specification used to define and execute the statistical model.

```
#form_arr = c(independendVar, usercovVar,covfixedVar, smoothVar, logVar, interactionVar, sqVar, sqVar_SQUA

### similarly for the independent variable...
#if independent variable is a smooth variable, log variable, or squared variable remove independendendVar
if(independendVar %in% c(  substring(smoothVar,3,nchar(smoothVar)-1),
                           smoothVarInt.stripped.term1,
                           substring(logVar,5,nchar(logVar)-1),
                           sqVar ) ){
  form_arr = c(usercovVar,covfixedVar, smoothVar.all, logVar, interactionVar, sqVar, sqVar_SQUARED);
  #form_arr = form_arr[form_arr != independendVar]
}
form_arr = form_arr[form_arr!=""]
#take out duplicate variables
form_arr = form_arr[!duplicated(form_arr)]

formulastr = paste(dependendVar," ~ ",paste(form_arr,collapse='+'))
#if(length(smoothVar) > 0){
#  formulastr = paste(dependendVar," ~ ",paste(form_arr,collapse='+'),"+",smoothVar)
#}

#get variables involve in the formula
varList = all.vars(as.formula(formulastr));
varList.independent = varList[-1]
print(varList)
print(formulastr)

########################
##  data manipulation  ##
########################
#data = data[data$eventname == "baseline_year_1_arm_1",]
print(summary(data[[independendVar]]))
#if independent variable has 5 or less unique values change it to character/factor variable
categorical.independent = FALSE
#if( length(table(data[[independendVar]])) < 6 ){
#  data[[independendVar]] = as.character(data[[independendVar]])
#  categorical.independent = TRUE
#} else{
#  data[[independendVar]] = as.numeric(as.character(data[[independendVar]]))
#}

if(class(data[[independendVar]]) != "numeric"){
   categorical.independent = TRUE
}

#user defined covariates
#for(ucov in unlist(inputs[['usercov.']])){
#  data[[ucov]] = as.numeric(as.character(data[[ucov]]))
#}

print(summary(data[[independendVar]]))
#determine if logistic regression or not
categorical.dependent = FALSE
data[[dependendVar]][data[[dependendVar]] == ""] = NA
if( length(table(data[[dependendVar]])) == 2 ){
  data[[dependendVar]] = as.factor(data[[dependendVar]])
  categorical.dependent = TRUE
} else{
  data[[dependendVar]] = as.numeric(as.character(data[[dependendVar]]))
}
#data[[dependendVar]] = as.numeric(as.character(data[[dependendVar]]))

#if("demo_prnt_marital_v2" %in% colnames(data)){
```

Dependent Variable (Y-axis)

nihtbx_fluidcomp_uncorrected

Independent Variable (X-axis, for plotting)

nihtbx_picvocab_uncorrected

Grouping Variable (for interaction and plotting)

Select subset of sessions

Other Independent Variables

Fixed Effect Covariates

Race/Ethnicity | SEX | EDU | Income | Marital | AGE

Random Effects

FAMILY | SITE

Expert Mode (testing and debugging)

submit

This application fits generalized additive mixed models using the R package GAMM4 (Simon Wood, Fabian Scheipl). The GAMM model is appropriate for both cross-sectional and longitudinal regression analyses and allows for an explicit modelling of aspects of the study design such as nesting of subjects within sites and family, twin pairs and other siblings.

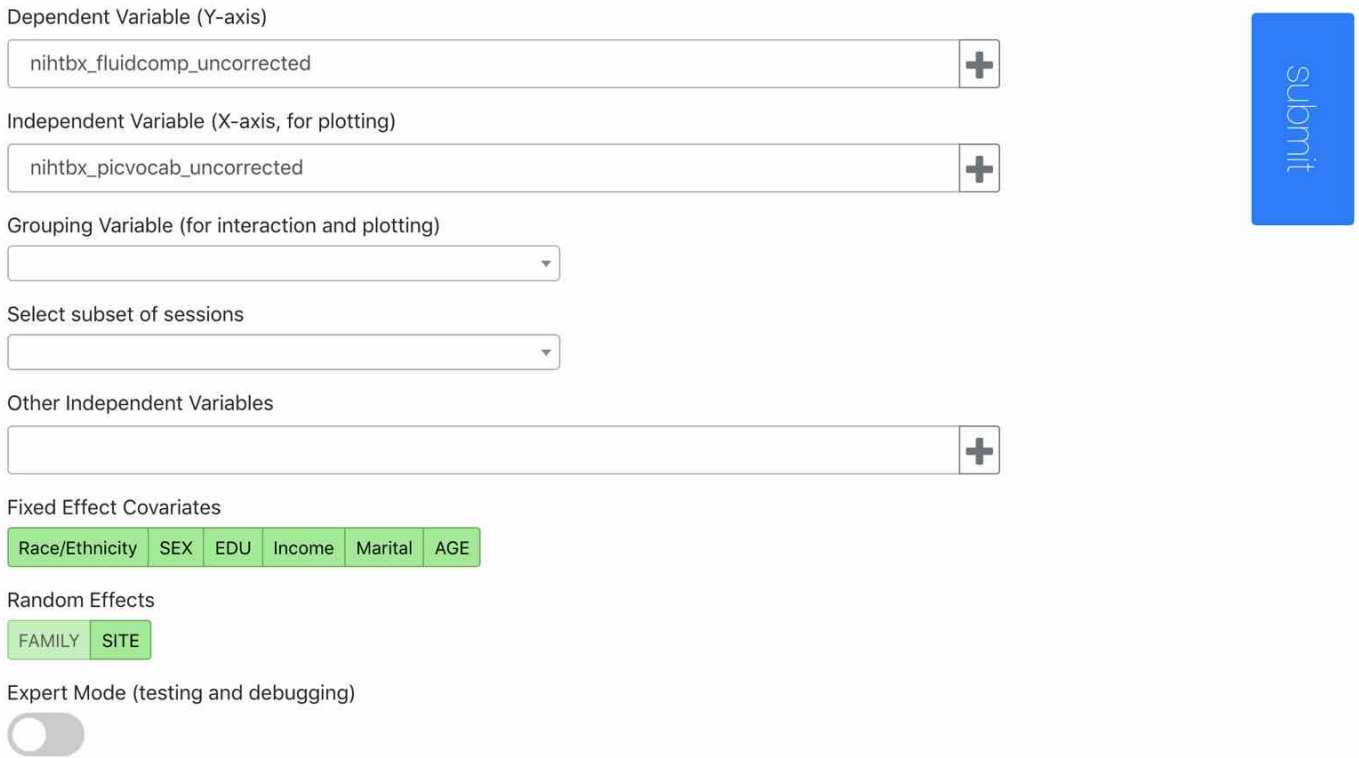
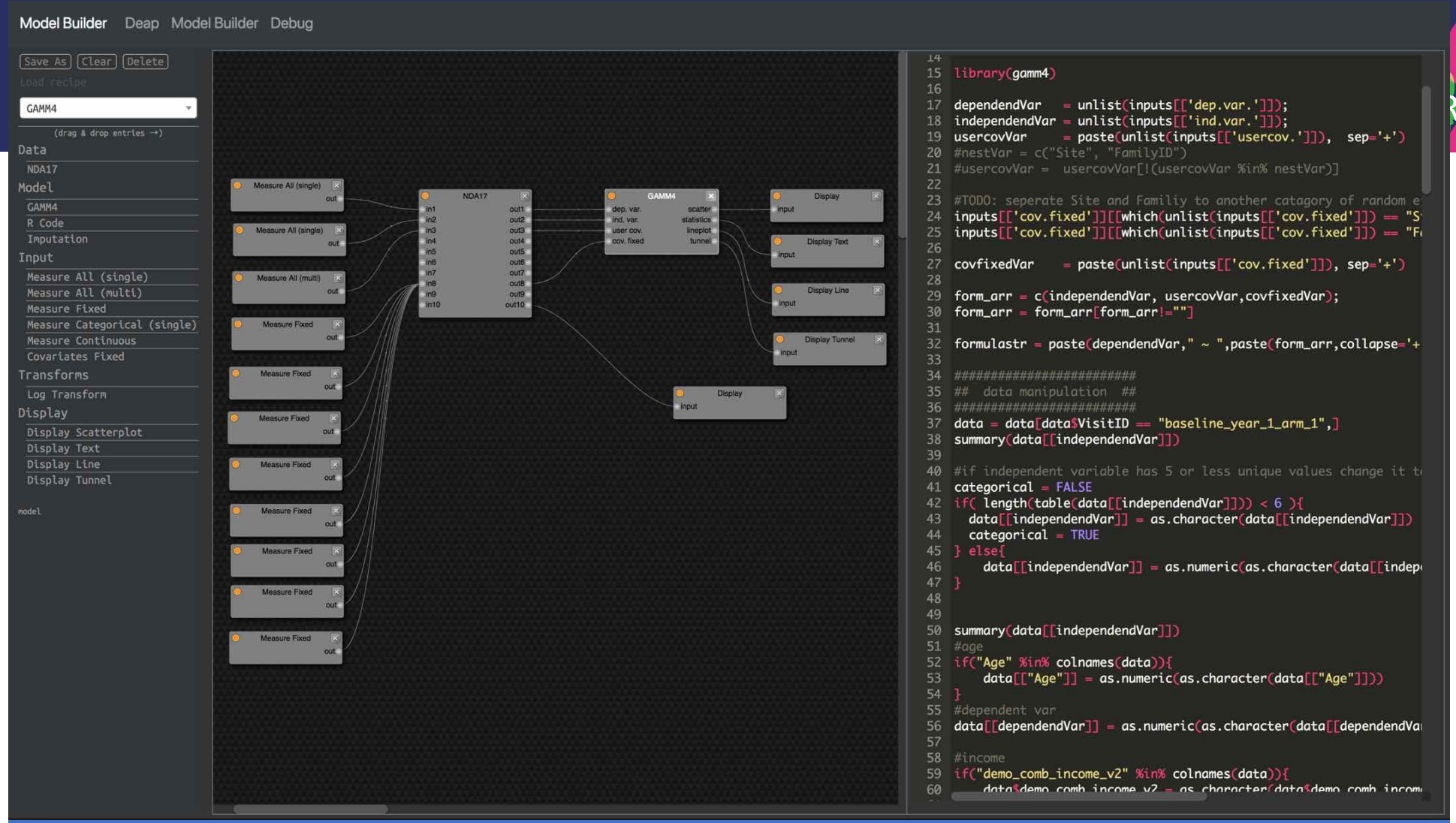


Fig. 1: Model specification used to define and execute the statistical model.

`10.17sec for calculation`

The model specified in Fig. 1 is used to estimate the statistical relationship between an independent variable and a measured dependent variable. In the generated model plot (Fig. 3) the independent variable is displayed on the X-axis and the dependent variable appears on the Y-axis. Both measures are user defined and can be selected from a list of available measures. Whereas the independent variable can be of any type (categorical or continuous), and there are no



Advanced Usage (Model Builder)

A collaborative environment to integrate advanced statistical analysis features into ABCD. The model builder is software agnostic. R modules coexist next to python/pandas, Matlab. Data frames are used for inter-nodal communication. System provides computational cloud resources and each block can be extracted from the system (data and source-code) for documentation and offline analysis.

DEAP Updates

- Population Weighting
- Image Analyses
- Enhanced interactive download of data
- Longitudinal analyses
- Twin analyses
- Cross-validation / out-of-sample estimation
- Machine Learning
- Missing data imputation