A data pipeline for studies on substance use and brain differences in the UK Biobank

This data processing pipeline consists of a set of scripts to be run in order to accomplish the steps of preparing the data for use in regression models to study the association between brain differences and alcohol and smoking phenotypes.

This pipeline is developed and running on R version 4.2.1 (Windows).

Variables to set at the top of each script:

Run date = date the pipeline is being run as a string formatted yyyy-mm-dd

Input path = string filepath where UKB data sets are stored

Output path = string filepath where pipeline script outputs are stored.

* Note that the file locations for inputs and outputs should be appropriate for storing UKB data. Up until the last stages of the pipeline outputs still contain protected UKB data. Do NOT set the output path as the repo.

R packages used in the pipeline

* dplyr (1.0.9)
* tidyr (1.2.0)
* reshape (0.8.9)
* broom (1.0.0)
* mice (3.14.0)

Packages used for visualizing various data checks through the pipeline

* ggplot2 (3.4.0)
* kableExtra (1.3.4)
* viridis (0.6.2)
* forcats (0.5.1)

**Pipeline step 0 IDP names**

UKB performs MRI data collection and analysis (Alfaro-Almagro et al., 2018; Miller et al., 2016). Rather than distributing the full MRI data, the UKB provides tabulated, MRI derived summary measures of brain structure and function called Imaging Derived Phenotypes (IDPs). Each IDP has a data field number and name indicating what brain feature it represents. These can be viewed on the UKB data showcase (<https://biobank.ndph.ox.ac.uk/showcase/label.cgi?id=100>).

In this pipeline script we import a text file (IDPs\_37741\_names.txt, provided in this repo) with the IDP data field numbers and names and convert this to a table which can be joined to other data sets by IDP data field number. We add additional columns specifying the MRI sequence, parcellation map, and other information. This allows sorting of a data set of IDPs by sequence, etc.

Inputs:

* IDPs\_37741\_names.txt: text file with IDP data fields and names (provided in repo)
* Yoonhoo\_assigned\_IDP\_regions\_11-30-21.csv: spreadsheet of IDPs with regions assigned by hand (provided in repo)

Columns in the IDP names table:

* IDP: UK Biobank data field for the IDP (ours are formatted X25009.2.0)
* IDP\_name: Human readable name of what the IDP represents
* numeric\_IDP: just the numeric part of the IDP data field (ex. 25009 instead of X25009.2.0)
* MRI: MRI sequence the IDP is derived from (ex. T1, T2, dMRI, etc.)
* map: which parcellation scheme the IDP is part of (ex. FAST, FIRST)
* dimension: area, thickness, or volume (relevant for T1 IDPs, NA for others)
* dMRI: Which NODDI measure the IDP represents (FA, MD, ICVF, ISOVF, MO, OD, L1, L2, L3, relevant only to dMRI, NA for other IDPs)
* region: what region the IDP is in (determined by hand, credit to Yoonhoo Chang for this)
* cortical: is the IDP cortical or subcortical (determined by hand)
* hemisphere: right or left hemisphere (or neither)
* paired: if the IDP is part of a left / right pair, this column contains the corresponding IDP on the opposite hemisphere
* display\_group: a tag useful for making manhattan plots of the IDPs

Derived IDPs:

The IDP name table includes some IDP names that are not provided by the UKB. For certain analyses, we included what I called “derived IDPs”. These are measures calculated by adding up or otherwise combining UKB IDPs. For example, we obtain a measure of total white matter volume from the Freesurfer ASEG parcellation by adding IDP X26553.2.0 “Volume of CerebralWhiteMatter (left hemisphere)” and IDP X26584.2.0 “Volume of CerebralWhiteMatter (right hemisphere)”. If you do not need these you can simply disregard them. If you want to derive your own IDPs, just modify the code in the derived IDPs section.

rfMRI ICs

We also include names for the 6 independent components (ICs) derived from rfMRI per the procedure described by Elliot et al (Elliott et al., 2018). This table does not include other rfMRI IDPs.

Errata (2024-05-01)

Four IDPs from the Freesurfer ASEG parcellation come out with the wrong dimension.

* X26548.2.0 Mean intensity of Accumbens-area (left hemisphere)
* X26564.2.0 Volume of Accumbens-area (left hemisphere)
* X26579.2.0 Mean intensity of Accumbens-area (right hemisphere)
* X26595.2.0 Volume of Accumbens-area (right hemisphere)

Temp fix for this

```{r}

IDP\_name\_table <- IDP\_name\_table %>%

mutate(dimension = ifelse(IDP %in% c("X26564.2.0", "X26595.2.0"), "volume", dimension))

```

Output

At the end of the script the IDP name table with the expanded information that has been added is saved as an RData file and as a .csv in the output path.

* IDP\_names\_table\_yyyy-mm-dd.RData
* IDP\_names\_table\_yyyy-mm-dd.csv

**Pipeline step 1 participants imaging**

In this step we import the data set of IDPs, join it to the phenotype data by participant ID, and do inclusion / exclusion on the imaging cohort.

Inputs

* UKB IDP file: .csv file with participant IDs in the first column. Each row is a participant, each column is an IDP.
  + Check the formatting of your IDP column names, this script assumes X25009.2.0 (example)
* subs\_netica.txt: output from rfMRI independent components analysis (ICA). See Elliot et al 2018 for this (Elliott et al., 2018), or drop this if you aren’t using rfMRI independent components (ICs).
* ID\_key\_for\_my\_UKB.txt: text file to match participant IDs in the imaging data to participant IDs in the phenotype data (you may not need to do this)
* UKB\_withdrawn.csv: .csv file from UKB with list of participant IDs for those who have requested to withdraw from the study.
* UKB relatedness matrix: file from UKB with kinship coefficient for each pair of participants

Main steps:

* Import the imaging data
* Ensure participant IDs in the imaging data match IDs in the phenotype data so they can be joined
  + If these already match in your data set you can delete the chunks about matching IDs
* Filter out withdrawn participants (IMPORTANT!)
* Remove participants with neurological disease
* Filter so that no two participants are greater than 3rd degree relatives
* Split IDPs into separate data sets by MRI sequence / processing / parcellation
* Normalize IDPs using z-score

Neurological disease:

UKB data field 20002 (https://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=20002) provides codes for self-reported non-cancer illnesses. Participants were asked in the touchscreen survey if a doctor had ever told them they have a serious illness. If they responded yes, they were asked about the illness during an interview with a trained nurse who assigned the illness the correct code (https://biobank.ndph.ox.ac.uk/showcase/refer.cgi?id=100235). If there was any uncertainty as to the illness, the participant described it to the nurse who used their expertise to code it correctly. Illnesses are coded with data coding 6 (https://biobank.ndph.ox.ac.uk/showcase/coding.cgi?id=6). If the illness could not be determined well enough to code it during the interview the nurse entered a free text description which was later coded by a physician. If the illness could not be determined after this procedure it was coded as 99999 “unclassifiable”.

We referenced the codes in Data Field 20002 to exclude participants with certain neurological conditions. We based our list of conditions to exclude on the exclusion criteria in Gray 2020 (Gray et al., 2020). The number of participants ultimately excluded for neurological conditions is slightly less than the sum of the numbers with neurological conditions as some participants have multiple conditions. Of importance for this study, we excluded participants with the condition dementia / Alzherimers / cognitive impairment (code 1263 in Data Coding 6).

Relatedness

Within the UK Biobank a number of participants share a familial relationship. We identified relatives of 3rd degree or closer using a kinship matrix provided by the UK Biobank. Per Bycroft 2018, 1st degree relatives have an expected kinship coefficient of ¼, 2nd degree relatives have an expected kinship coefficient of 1/8, while 3rd degree relatives have a kinship coefficient of 1/16 (Bycroft et al., 2018). In order to ensure all 3rd degree or closer relatives were identified we first filtered for kinship coefficients equal to or greater than 1/20. Within that group we identified related pairs with both individuals in the neuroimaging set. Within those pairs, we implemented an algorithm based on the one developed by Hanscombe et al (Hanscombe et al., 2019). In this algorithm we dropped participants in the order of the most highly interconnected until we were left with only those individuals who are connected to one other person. We then randomly selected on member of each pair to remove.

Splitting the IDPs

In this script we split the IDPs up into several RData files. There are several reasons for this:

1. The main IDP data set is very large, processing can be sped up by splitting into smaller files.
2. The rate of missing data varies by MRI sequence. By separating IDPs derived from T1, T2, and dMRI, participants who are missing one sequence can still be retained for the other sequences. The n tracker output by this script gives the n for each sequence.
3. In this version of the pipeline we retain all IDPs for further processing. However, it may not make sense to keep every single IDP for a given analysis. A number of the sets represent the same regions under different parcellation schemes (Smith et al., 2020) and it is a duplication to include both in the same analysis.
   1. I suggest filtering IDPs to only those you plan to analyze in pipeline step 1 before going on to future steps. The IDP names table can be used for filtering.

Outputs

* n\_tracker\_p1: RData file with number of participants in the imaging cohort by each step of processing and inclusion / exclusion.
* neuro\_disease\_table: RData file with data frame containing the neurological disease name, code from data field 20002, and number of participants with that code.
* total\_volumes: IDPs representing measures of total brain volume
  + This set is NOT z-score normalized
* control\_measures: normalized IDPs which may be used as controls, for example controlling for total brain volume or average cortical thickness
  + There is some redundancy with total\_volumes
* GMV: normalized IDPs from the T1 FAST and FIRST parcellations
* free\_2009: normalized IDPs from the T1 Freesurfer a2009s parcellation
* desikan: normalized IDPs from the Desikan Killiany parcellation
* T1\_regional: normalized IDPs from Freesurfer ASEG and Freesurfer subseg
* T2: normalized IDPs derived from T2\* MRI
* dMRI: normalized IDPs derived from dMRI
* rfMRI: normalized rfMRI ICs
* filtered\_eid\_p1: list of participant IDs which are included in at least one MRI sequence and should be included in processing of phenotype measures

**Pipeline step 2 imaging covariates**

Alfaro Almagro et al describe in detail controls to be included with all analyses of the UKB imaging data to address potential confounding (Alfaro-Almagro et al., 2021). In this script we implement the recommended procedure.

Inputs

* IDP\_imaging\_covariates.csv
  + Data fields:
    - Participant ID
    - X31.0.0 (sex, registry and corrected in touchscreen)
    - X53.2.0 (imaging appointment data)
    - X54.2.0 (imaging site)
    - X21003.2.0 (age at imaging)
    - X25000.2.0 (head size)
    - X25741.2.0 (rfMRI motion)
* n\_tracker\_p1: RData file with number of participants in the imaging cohort by each step of processing and inclusion / exclusion (saved in pipeline step 1)
* filtered\_eid\_p1: list of participant IDs which are included in at least one MRI sequence and should be included in processing of phenotype measures (saved in pipeline step 1)

Main steps

* Import the imaging covariate data
* Ensure participant IDs match the phenotypes and the filtered eid from pipeline step 1
* Capture the number participants missing data for the covariates
  + Participants are only dropped if they are missing head size
* Scale the covariates using median and median absolute deviation (MAD) \* 1.48
* Remove outliers greater than 8
* Split the covariates on site
* Scale the split covariates using mean and sd (z-score normalization)
* Add non-linear terms

Detailed description:

We processed the imaging confounds in accordance with the procedure described by Alfaro Almagro 2021. Processing was done on the subset of participants who were included in any of the IDP subsets. For the imaging confounds we included sex (31.0), imaging site (54.2), date of attending imaging site (53.2), date^2, age at imaging (21003.2), age^2, age\*sex, head size (25000.2), and rfMRI motion (25741.2). We decided not to include tfMRI motion as it has a high rate of missingness, is less reliable than rfMRI, and is redundant.

Date of attending the imaging site was converted into a numeric as days since Jan 1 1970 (https://statistics.berkeley.edu/computing/faqs/dates-and-times-r). All imaging confounds except sex and imaging site were then demeaned and normalized using the median and median absolute deviation \* 1.48 (one standard deviation). After demeaning all outliers greater than 8 were removed and replaced with NAN. Since site is considered the most influential confound (Alfaro-Almagro et al., 2021) the confounds were split on site. Within each site, missing values and outliers were replaced with the site median. All the imaging confounds had a rate of missing of 0.02% or less. The values for each site (except the site value itself) were then normalized to a z-score within that site. Sex was transformed into a z-score within each site in order to facilitate creating a combined variable with age. Non-linear confounds were modelled by including date^2, age^2, and age \* sex. The per site confounds were then renamed to create unique variables for each site (ex. site 1 age). These variables hold the values for their corresponding site and zeros for the other two sites. This resulted in 25 covariates which were included in our regression analysis.

Outputs

* Imaging\_covariates: RData file with imaging covariates that have not been processed, these are used for imputation and demographics later in the pipeline.
* processed\_imaging\_covariates: RData file with imaging covariates which are split on site and normalized according to procedure in Alfaro Almagro 2018 (Alfaro-Almagro et al., 2021). These will be included in regression analyses of neuroimaging data
* n\_tracker\_p2: RData file with number of participants in the imaging cohort by each step of processing and inclusion / exclusion.
* filtered\_eid\_p2: list of participant IDs which are included in at least one MRI sequence and should be included in processing of phenotype measures, less any participants removed for missing head size in pipeline step 2
* missing\_report\_p2: rate of missing data in the imaging covariates

**Pipeline step 3 smoking**

Set imaging vs. baseline:

Set the touchscreen variable to “imaging”, run the script, then set “baseline” and run it again. This results in two outputs, one for the imaging visit and the other for baseline.

Participants in the imaging cohort were administered the touchscreen survey with questions about smoking and alcohol use on two occasions, the baseline appointment and again when they returned for neuroimaging. This script runs on both the baseline or imaging smoking data. Our approach is to take the imaging responses as the primary variable but if data is missing for the imaging appointment, we backfill it with data from the baseline appointment if that is available. This reduces the number of values which are missing and need to be imputed in pipeline step 5.

Inputs

* filtered\_eid\_p2: list of participant IDs which are included in at least one MRI sequence and should be included in processing of phenotype measures (saved in pipeline step 2)
* missing\_report\_p2: rate of missing values from covariates handled in prior pipeline steps
* smoking data fields from imaging visit as .csv
  + Data fields:
    - Participant ID
    - n\_1239\_2\_0 Current tobacco smoking
    - n\_1249\_2\_0 Past tobacco smoking
    - n\_2644\_2\_0 Light smokers
    - n\_20161\_2\_0 Pack years
    - n\_3456\_2\_0 Number of cigarettes currently smoked daily
    - n\_2887\_2\_0 Number of cigarettes previously smoked daily
  + Language note:
    - UKB does not always use person first language in their data field names (ex. Light smokers). I have given the data field names exactly as they appear in the showcase for clarity but want to flag this so you can choose person first language for any finished products.
* smoking data fields from baseline visit as .csv
  + Same fields as for imaging, but with \_0\_0 suffix

For smoking consumption, participants indicated current (Data field 1239, Current tobacco smoking) or past smoking (Data field 1249, Past tobacco smoking). Participants who endorsed daily or near daily smoking answered questions about age of smoking onset and offset and quantity of cigarettes smoked to calculate pack years and cigarettes per day. We assigned those who never smoked or smoked less than 100 cigarettes in their lifetime 0 pack years. For those who endorsed occasional smoking with greater than 100 cigarettes over their lifetime, but not daily smoking, we assigned 1 pack year history of smoking (yes in Data field 2644). Ever daily smoked is an indicator variable for those who endorsed daily or near daily smoking.

Never smoked is those who have smoked less than 100 cigarettes in their lifetime. We determined smoking status (Current, Former, Never) using data fields 1239, 1249, and 2644.

In some models we explored use of cigarettes per day instead of pack years to represent smoking since it is more parallel to the drinks per week measure. participants indicated current (Data field 1239, Current tobacco smoking) or past smoking (Data field 1249, Past tobacco smoking). Participants who endorsed daily or near daily smoking currently or in the past were asked about the number of cigarettes per day either currently (Data field 3457, Number of cigarettes currently smoked daily) or in the past (Data field 2887, Number of cigarettes previously smoked daily). We combined these into a single measure of cigarettes per day (current and past). Those who never smoked or smoked less than 100 cigarettes in their lifetime (“No” in Data field 2644 Light smokers) were assigned 0 pack years. For those who endorsed occasional smoking with greater than 100 cigarettes over their lifetime, but not daily smoking, we assigned 1 cigarette per day (“Yes” in Data field 2644).

Outputs

* imaging\_smoking: RData file with processed smoking data fields
  + Participant ID
  + pack\_years
  + cig\_per\_day
  + ever\_daily\_smoked
  + never\_smoked
  + smoking\_status
* Run the script a second time after switching the touchscreen variable to get baseline smoking variables

**Pipeline step 4 alcohol**

Inputs

* filtered\_eid\_p2: list of participant IDs which are included in at least one MRI sequence and should be included in processing of phenotype measures (saved in pipeline step 2)
* missing\_report\_p2: rate of missing values from covariates handled in prior pipeline steps
* alcohol\_imaging: alcohol data fields from the imaging visit as a .csv
  + Data fields:
    - n\_1558\_2\_0 Alcohol intake frequency
    - Number of drinks per week in those drinking on a weekly basis
      * n\_1568, n\_1578, n\_1588, n\_1598, n\_1608, n\_5364
    - Number of drinks per month in those drinking less than weekly
      * n\_4407, n\_4418, n\_4429, n\_4440, n\_4451, n\_4462
    - n\_3731\_2\_0 Former alcohol drinker (ditto note about non-person first language)
* alcohol\_baseline: alcohol data fields from the baseline visit as a .csv
  + Same fields as for imaging, but with \_0\_0 suffix

Deriving the week drinks measure

In the touchscreen survey administered at the baseline and imaging appointments, participants indicated alcohol use frequency (data field 1558 Alcohol intake frequency). Participants who endorsed drinking were asked to estimate how much they drink in a typical week (for those who drink on a daily to weekly basis) or in a typical month (for those who drink monthly or less) in units of red wine, white wine, fortified wine, beer and cider, spirits, or other (such as alcopops). If participants’ intake varied, participants were instructed to estimate their average intake over the past year (https://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=1558). To derive a standard measure of drinks per week, we summed alcohol consumption across different drink types to get total units of drinks consumed per week or month. For those who provided a monthly estimate we converted the monthly totals to weekly estimates by dividing the amount by 4.3.

The variable never alcohol distinguishes participants who indicated no alcohol consumption in data field 1558 Alcohol intake frequency and that they never consumed alcohol in data field 3731 Former alcohol drinker (this is UKB data field name and is not person first). Since both those who have never consumed alcohol and those who formerly consumed alcohol it can be useful to distinguish. The group of participants who formerly consumed alcohol may contain people in remission from AUD and who have very high lifetime alcohol exposure despite current non-consumption.

Outputs

* imaging\_alcohol: RData file with alcohol covariates
  + Participant ID
  + drink\_freq
  + week\_drinks
  + never\_alcohol
  + alcohol\_status
* missing\_report\_p4: RData file with rates of missing data in alcohol fields as well as prior pipeline steps

Pipeline step 5 covariates, backfilling, and imputation

In this script we bring in and do some processing on covariates which aren’t related to the neuroimaging, alcohol, or smoking. If there is a value from the imaging visit, we take that value. If it is missing, we check if there is a value from the baseline visit and take that. This reduces the rate of missing data which has to be imputed. Finally, we impute the data points which are missing from both the imaging and baseline visits using multiple chained equations (MICE) with the classification and regression trees method (CART) (van Buuren, 2018; van Buuren & Groothuis-Oudshoorn, 2011). Before imputation we z-score normalize the numeric variables. After imputation, we keep one version with normalized values for use with the regression models. We also use the saved means and standard deviations to create a second version which is has missing values imputed but is not normalized for use in the demographics table.

Inputs

* Files from prior pipeline steps:
  + filtered\_eid\_p2: list of participant IDs which are included in at least one MRI sequence and should be included in processing of phenotype measures (saved in pipeline step
  + imaging\_covariates: imaging covriates from pipeline step 2, these are the set that is NOT normalized and split on site. They are included to increase the accuracy of MICE.
  + missing\_report\_p4: rate of missing values from covariates handled in prior pipeline steps
  + baseline\_smoking: output from pipeline step 3, smoking variables derived from baseline visit values
  + imaging\_smoking: output from pipeline step 3, smoking variables derived from imaging visit values
  + imaging\_alcohol: output from pipeline step 4, alcohol variables from the imaging visit and backfilled from the baseline visit. The rate of missing values before and after backfilling is given in the missing report
* UKB data fields, both imaging visit values (suffix \_2\_0) and baseline (suffix \_0\_0)
  + - Note I have these divided into various .csv files, the reasons for this are entirely historical. Can be nice if you have any limitations on file size.
  + 2178 Health rating
  + 2188 Long-standing illness, disability or infirmity
  + 2443 Diabetes
  + 4056 Age at first stroke
  + 20002 Non-cancer illness code
  + 21001 BMI
  + 738 Income
  + 21002 Weight
  + 25009 Brain volume
  + 6145 Stress
  + 884 Times per week moderate exercise
  + 904 Times per week vigorous exercise
  + 2453 Cancer
  + 2473 Other diagnosis

Detailed info on covariates and their processing prior to imputation:

Diastolic and systolic blood pressure (BP), data fields 4079, 94, 4080, 93

Diastolic BP from the baseline and imaging visits is recorded in data field 4079, and systolic BP in data field 4080. This reflects a reading obtained with an automatic blood pressure cuff. When the automatic blood pressure cuff fails to get a reading the blood pressure was taken manually and recorded in field 94 (diastolic) and 93 (systolic) (https://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=94). If the automatic reading is missing, we replaced it with the manual value if available for both the imaging and baseline values. If both automatic and manual readings were missing from the imaging visit, we replaced with BP from the baseline visit if available.

Body mass index (BMI), data field 21001

Participant height (data field 50) and weight (data field 21002) were collected during the visit and used to calculate BMI (data field 21001). If BMI values are missing for the imaging visit we backfilled with values from the baseline visit.

Waist Hip Ratio, derived

Participant waist (data field 48) and hip circumference (data field 49) were collected during the visit. We divided the waist measurement by the hip measurement to obtain a combined measure (Gray et al., 2020). We calculated waist hip ratio for the imaging visit values and baseline visit values separately. If the imaging visit is missing either waist, hip, or both, we backfilled with the baseline visit value if available.

Income, data field 738

Income is recorded in data field 738 “total household income before taxes”. Participants indicated which income band they fell into, < 18000, 18000 to 30999, 31000 to 52000, 52000 to 100000, and greater than 100000. We converted this into a continuous numeric value by taking the lower bound of the income band. We performed the conversion on both the imaging visit and baseline values. We took the imaging visit value as primary and if the imaging value is missing we replaced it with the baseline value. Responses of “Do not know” and “Prefer not to answer” are treated as missing.

Educational Attainment (Qualifications), data field 6138

We referenced data field 6138 “Which of the following qualifications do you have” to determine education level. We converted this into numeric years of education using the conversion system developed by Zhou et al and based on the International Standard Classification of Education (Zhou et al., 2021). We performed the conversion on both the imaging visit and baseline values. We took the imaging visit value as primary and if the imaging value is missing we replaced it with the baseline value. Responses of “None of the above” and “Prefer not to answer” are treated as missing.

Stress, data field 6145

Participants were asked to indicate if they had experienced any of a list of stressors within the past 2 years, with the option to select more than 1. The stressors include “serious illness, injury, or assault of the participant or a close relative”, “death of a close relative, spouse, or partner”, “marital separation or divorce”, “financial difficulties”, and “none of the above” (<https://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=6145>). These responses are recorded in data field 6145 with data coding 100502. We converted the data coding to a binary variable with 1 indicating the participant had experienced any of the stressors within the past 2 years. This conversion was performed on both the imaging visit and baseline values. We took the imaging visit value as primary and if the imaging value is missing we replaced it with the baseline value. Responses of “Prefer not to answer” are treated as missing.

Moderate and vigorous exercise, data fields 884, 904

Participants indicated the number of days per week (0 to 7) on which they performed at least 10 minutes pf moderate or vigorous physical activity. The number of days performing moderate physical activity is provided in data field 884 and the number of days performing vigorous physical activity is provided in in data field 904. We took the imaging visit value as primary and if the imaging value is missing we replaced it with the baseline value.

History of stroke, data field 4056

Participants were first asked if they had been diagnosed with a vascular or heart problem by a doctor "Has a doctor ever told you that you have had any of the following conditions?”. Those who indicated stroke were then asked their age when they received the diagnosis and this is recorded in data field 4056 “Age stroke diagnosed”. If participants have a value in this data field we consider them to have a positive stroke history. This approach is based on (Cole, 2020). We encode this as a binary variable with 1 = history of stroke. There are very few participants with stroke history in our cohort due to the exclusion of participants with a history of neurological disease. We took the imaging visit value as primary and if the imaging value is missing we replaced it with the baseline value.

History of diabetes, data field 2443

Participants were asked if they had ever been told by a doctor that they have diabetes. We take those who indicated “Yes” as having a history of diabetes. We encode this as a binary variable with 1 = history of diabetes. We took the imaging visit value as primary and if the imaging value is missing we replaced it with the baseline value. Responses of “Do not know” and “Prefer not to answer” are treated as missing values.

History of cancer, data field 2453

Participants were asked if they had ever been told by a doctor that they have cancer. We take those who indicated “Yes” as having a history of cancer. We encode this as a binary variable with 1 = history of cancer. We took the imaging visit value as primary and if the imaging value is missing we replaced it with the baseline value. Responses of “Do not know” and “Prefer not to answer” are treated as missing values.

History of other serious medical condition / disability diagnosed by a doctor, data field 2473

Participants were asked if they had ever been told by a doctor that they have “any other” serious medical condition or disability. This prompt comes after the questions about vascular and heart disease, diabetes, and cancer. Those who indicated “Yes” are considered to have a history of serious medical condition or disability. We encode this as a binary variable with 1 = history of a serious condition or disability. We took the imaging visit value as primary and if the imaging value is missing we replaced it with the baseline value. Responses of “Do not know” and “Prefer not to answer” are treated as missing values.

Health rating, data field 2178

Participants were asked "In general how would you rate your overall health?” Categorical responses are encoded using data coding 100508. We retained the numeric scale from this data coding corresponding to levels of perceived health. We took the imaging visit value as primary and if the imaging value is missing we replaced it with the baseline value. Responses of “Do not know” and “Prefer not to answer” are treated as missing values.

Data coding 100508:

1 Excellent

2 Good

3 Fair

4 Poor

-1 Do not know

-3 Prefer not to answer

Longstanding illness, data field 2188

Participants were asked “Do you have any long-standing illness, disability or infirmity?". Those who indicated “Yes” are considered to have a longstanding illness or disability. We encode this as a binary variable with 1 = presence of a longstanding condition. We took the imaging visit value as primary and if the imaging value is missing we replaced it with the baseline value. Responses of “Do not know” and “Prefer not to answer” are treated as missing values.

Psychiatric

In our model we wanted to control for the presence of psychiatric conditions. We referenced Data Field 20002. UKB data field 20002 (https://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=20002) provides codes for self-reported non-cancer illnesses. Participants were asked in the touchscreen survey if a doctor had ever told them they have a serious illness. If they responded yes, they were asked about the illness during an interview with a trained nurse who assigned the illness the correct code (https://biobank.ndph.ox.ac.uk/showcase/refer.cgi?id=100235). If there was any uncertainty as to the illness, the participant described it to the nurse who used their expertise to code it correctly. Illnesses are coded with data coding 6 (https://biobank.ndph.ox.ac.uk/showcase/coding.cgi?id=6). If the illness could not be determined well enough to code it during the interview the nurse entered a free text description which was later coded by a physician. If the illness could not be determined after this procedure it was coded as 99999 “unclassifiable”.

We reference this data field earlier in data processing to identify participants with neurological conditions which could influence their brain imaging and exclude them from the sample. For psychiatric conditions, we use the data field to create an indicator variable for participants with any of a set of psychiatric conditions. The participants remain in the study cohort and we include the variable “psychiatric” in our regression analyses. The psychiatric illnesses we incorporate into this variable are depression (code 1286), anxiety / panic attacks (code 1287), nervous breakdown (code 1288), schizophrenia (code 1289) and manic / bipolar disorder / manic depression (code 1291). These illness, codes, and the number of participants with each are reported in supplemental table X. The number of participants coded as having a psychiatric illness is slightly less than the sum of the numbers for each individual illness as some participants have multiple diagnoses. UKB also makes available the results of questionnaires for psychiatric symptoms given as part of the touchscreen survey. We chose to focus on conditions which are diagnosed by a doctor rather than attempting to determine diagnosis from reported symptoms.

Outputs

* psych\_disease\_table: table with psychological illness, UKB disease code, and number of participants affected. Generated while creating the psych disease variable, these participants are not filtered out but the variable controls for any effects of psychiatric illnesses.
* pre-imputation: .csv with normalized values before imputation, this is done so you can re-import it and resolve a problem with the column data types
* post\_imputation: output from mice function without any further processing
* imputation\_results: normalized variables after imputing all missing values
* un\_scaled: variables after imputing all missing values, with normalization reverted
* missing\_report\_p5: rate of missingness in all variables before and after backfilling the imaging visit values with baseline

**Pipeline step 6 genetic principal components**

This script is simple, import genetic principal components, filter for participants in our imaging cohort, then select the first 10 principal components. We decided the first 10 are sufficient after consultation with experts.

* We do not have any participants in our cohort who have imaging data and are missing genetic data. If you do, you may want to move this script earlier in the pipeline and drop participants with no genetic data.

Inputs

* filtered\_eid\_p2: list of participant IDs which are included in at least one MRI sequence and should be included in processing of phenotype measures (saved in pipeline step 2)
* ukb\_pc20: .csv file of UKB calculated genetic principal components for each participant

Outputs

* genetic\_PC: RData file with first 10 genetic PCs for imaging cohort

**Pipeline step 7 regression**

Finally we combine all these processed variable sets into the regression model! I have functionalized the regression so you can pass the variables you want to use as strings, since not all variables get used in all versions of the model. In this version of the script I am running the regression on a large set of the IDPs and including controls for total brain volume / total thickness. In most specific analyses we run regression on a selected subset of IDPs and do not necessarily control for total volume, however, it’s easier to have this and delete it than add it if you do want that.

Inputs

* IDP\_names\_table: table with the IDP names (should match columns in the imaging data files) plus expanded information which makes it easier to sort and plot IDP data
* filtered\_eid\_p2: list of participant IDs which are included in at least one MRI sequence and should be included in processing of phenotype measures (saved in pipeline step 2)
* genetic\_PC: RData file with first 10 genetic PCs for imaging cohort
* processed\_imaging\_covariates: imaging covariates from pipeline step 2, these are the set that is IS normalized and split on site for inclusion with the regression model
* imputed\_covariates: normalized variables after imputing all missing values

Main steps:

* Import all RData files with processed variables and combine them
* Define the regression function
  + Function takes list of string names of variables to use as predictors, total volume control, and covariates
  + Use lapply to run glm model on a set of IDPs
  + Extracts beta, standard error (se) and p value for each variable for each IDP
  + Perform false discovery rate (FDR) correction on all p values using the p.adjust function (Benjamini & Hochberg, 1995)
  + Join the IDP names table with expanded information about the IDPs for sorting

Outputs

regression\_results: RData file with beta, se, and p value for all covariates as columns and IDPs as rows.

References

Alfaro-Almagro, F., Jenkinson, M., Bangerter, N. K., Andersson, J. L. R., Griffanti, L., Douaud, G., Sotiropoulos, S. N., Jbabdi, S., Hernandez-Fernandez, M., Vallee, E., Vidaurre, D., Webster, M., McCarthy, P., Rorden, C., Daducci, A., Alexander, D. C., Zhang, H., Dragonu, I., Matthews, P. M., … Smith, S. M. (2018). Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank. *NeuroImage*, *166*(April 2017), 400–424. https://doi.org/10.1016/j.neuroimage.2017.10.034

Alfaro-Almagro, F., McCarthy, P., Afyouni, S., Andersson, J. L. R., Bastiani, M., Miller, K. L., Nichols, T. E., & Smith, S. M. (2021). Confound modelling in UK Biobank brain imaging. *NeuroImage*, *224*(2021), 117002. https://doi.org/10.1016/j.neuroimage.2020.117002

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O’Connell, J., Cortes, A., Welsh, S., Young, A., Effingham, M., McVean, G., Leslie, S., Allen, N., Donnelly, P., & Marchini, J. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature*, *562*(7726), 203–209. https://doi.org/10.1038/s41586-018-0579-z

Cole, J. H. (2020). Multimodality neuroimaging brain-age in UK biobank: Relationship to biomedical, lifestyle, and cognitive factors. *Neurobiology of Aging*, *92*, 34–42. https://doi.org/10.1016/j.neurobiolaging.2020.03.014

Elliott, L. T., Sharp, K., Alfaro-Almagro, F., Shi, S., Miller, K. L., Douaud, G., Marchini, J., & Smith, S. M. (2018). Genome-wide association studies of brain imaging phenotypes in UK Biobank. *Nature*, *562*(7726), 210–216. https://doi.org/10.1038/s41586-018-0571-7

Gray, J. C., Thompson, M., Bachman, C., Owens, M. M., Murphy, M., & Palmer, R. (2020). Associations of cigarette smoking with gray and white matter in the UK Biobank. *Neuropsychopharmacology*, *45*(7), 1215–1222. https://doi.org/10.1038/s41386-020-0630-2

Hanscombe, K. B., Coleman, J. R. I., Traylor, M., & Lewis, C. M. (2019). ukbtools: An R package to manage and query UK Biobank data. *PLoS ONE*, *14*(5), e0214311. https://doi.org/10.1371/journal.pone.0214311

Miller, K. L., Alfaro-Almagro, F., Bangerter, N. K., Thomas, D. L., Yacoub, E., Xu, J., Bartsch, A. J., Jbabdi, S., Sotiropoulos, S. N., Andersson, J. L. R., Griffanti, L., Douaud, G., Okell, T. W., Weale, P., Dragonu, I., Garratt, S., Hudson, S., Collins, R., Jenkinson, M., … Smith, S. M. (2016). Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nature Neuroscience*, *19*(11), 1523–1536. https://doi.org/10.1038/nn.4393

Smith, S. M., Alfaro-almagro, F., & Miller, K. L. (2020). *UK Biobank Brain Imaging Documentation*. https://www.ukbiobank.ac.uk/

van Buuren, S. (2018). Classification and Regression Trees. In *Flexible Imputation of Missing Data*.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, *45*, 1–67. https://doi.org/10.18637/jss.v045.i03

Zhou, T., Sun, D., Li, X., Ma, H., Heianza, Y., & Qi, L. (2021). Educational attainment and drinking behaviors: Mendelian randomization study in UK Biobank. *Molecular Psychiatry*, *26*(8), 4355–4366. https://doi.org/10.1038/s41380-019-0596-9