

Normative modeling in Schizophrenia - Analysis of the 34 regions parcellation

Noemi González Lois

2021-06-03

Packages and libraries

```
library("easypackages") # install.packages("easypackages")

# get a list of all the needed packages
list.of.packages <- c("viridis", "tidyverse", "MatchIt", "grid", "png",
                     "gridExtra", "parallel", "nlme", "JMbayes",
                     "BiocManager", "Biostrings", "lme4", "ggplot2",
                     "Hmisc", "devtools", "longCombat", "neuroCombat",
                     "tinytex", "knitr", "dplyr")

# load them all
libraries(list.of.packages)
ncores <- detectCores() # Number of cores available in pc
rm(list.of.packages)
```

Set working directory and load functions

```
setwd("/data_J/Scripts")
source("1_DataPreparation.R")
source("2_RegressionModel.R")
source("3_Statistics.R")
source("4_EDA.R")
```

Data preparation

Options of DataPreparation function:

- parc = "parc35" or "parc308" (whether to use the 34 regions parcellation or the 308 regions one)
- harmonization = "lC" or "nC" (whether to use lonCombat or NeuroCombat harmonization)
- match = T or F (whether to use match-it or not)

```
df_lC_NO_matched <- DataPreparation(parc = "parc35",
                                   harmonization = "lC",
                                   match = F)
```

```
## File parc35 already exists, reading file...
## File with longCombat harmonization already exists, reading file...
## Done!
```

```
df_lC_matched <- DataPreparation(parac = "parc35",
                                harmonization = "lC",
                                match = T)
```

```
## File parc35 already exists, reading file...
## File with longCombat harmonization already exists, reading file...
## File with MATCH-IT already exists, reading file...
## Done!
```

Warning: Fewer control units than treated units; not all treated units will get a match.

Stack Overflow: This warning is because our treated group is larger than our control group (this happens in timepoint = 2). If you're doing 1:1 matching without replacement, all the control units will be used up before all the treated units get a match. **To remedy this, you need to match with replacement or think about whether you actually want to generalize to the control population and switch the labels on the treatment groups.** You can do this by creating a new variable, say notY, which is 1 - Y and then performing the same operations.

The match in **sex variable** is not exact. Anyway, the match in # patients, # controls is done well (is it enough?)

NO MATCHED DATASET

NO MATCHED	timepoint 1	timepoint 2	timepoint 3
# controls	298	293	109
# patients	169	168	50

Timepoint 1	sex 0	sex 1
# controls	131	167
# patients	38	131

Timepoint 2	sex 0	sex 1
# controls	130	163
# patients	38	130

Timepoint 3	sex 0	sex 1
# controls	50	59
# patients	7	43

MATCHED DATASET

Number of patients vs number of controls per timepoint is not exactly the same:

MATCHED	timepoint 1	timepoint 2	timepoint 3
# controls	169	164	49
# patients	169	164	49

Timepoint 1	sex 0	sex 1
# controls	38	131
# patients	38	131

Timepoint 2	sex 0	sex 1
# controls	37	127
# patients	38	126

Timepoint 3	sex 0	sex 1
# controls	8	41
# patients	7	42

Exploratory Data Analysis

Show relevant figures and analytics before and after data preparation. For example, age of controls vs age of patients in the raw df vs the match-it df:

Matching ages:

```
EDA_match_ages(df_1C_NO_matched,"NO match-it")
```

```
##
## Ages that doesn't match in patients vs controls in NO match-it dataset are:
## 69 68 67 65 64 63 62 60 59
```

```
EDA_match_ages(df_1C_matched,"match-it")
```

```
##
## Ages that doesn't match in patients vs controls in match-it dataset are:
## 60 58 57 56 46
```

Linear Mixed Effects Model Regression

Calling the `run_NormativeModel` function with different datasets. This will return the z scores (one for each region for each timepoint of each subject)

NO Match-it dataframe (longCombat):

```
Zs_NOMatch <- run_NormativeModel(df_lC_NO_matched,
                                measure = "CT_freesurfer",
                                parc = "parc35",
                                match = "NOMatch",
                                harmonization = "lC")
```

Match-it dataframe (longCombat):

```
Zs_match <- run_NormativeModel(df_lC_matched,
                                measure = "CT_freesurfer",
                                parc = "parc35",
                                match = "match",
                                harmonization = "lC")
```

Match-it dataframe (Age*Diagnosis) BEFORE EXCLUDING DEVIANTS

```
p_val <- run_AgeDiagnosisModel(df_lC_matched,
                               measure = "CT_freesurfer",
                               Z = NULL,
                               exclude_deviants = F)

p_val_FDR <- Apply_FDR_Correction(p_val)
```

Without FDR correction:

Variable dcode has statistical significance for 28 regions

Variable dcode_age has statistical significance for 14 regions

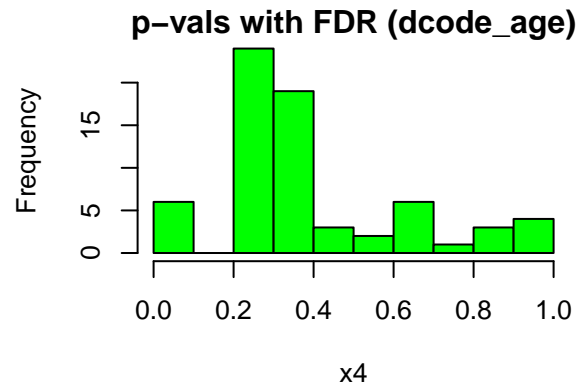
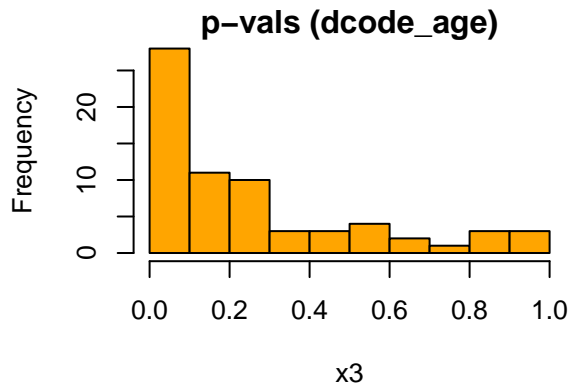
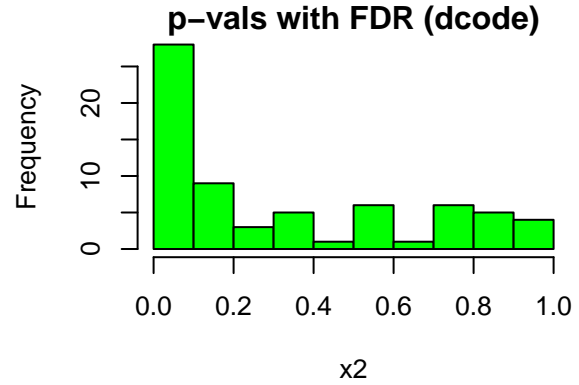
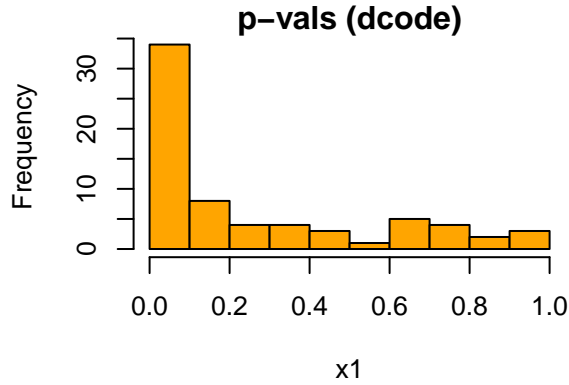
##	reg	scode	age	euler	dcode	dcode_age
## [1,]	0	0.902627774	1.376441e-10	1.585103e-03	0.05520999	0.311477261
## [2,]	0	0.002974282	4.919699e-02	6.363141e-02	0.69065034	0.000958305
## [3,]	0	0.011063046	0.000000e+00	1.653418e-02	0.14065095	0.242002275
## [4,]	0	0.656891364	1.284373e-06	3.036019e-05	0.47491087	0.813085773

With FDR correction:

Variable dcode has statistical significance for 20 regions

Variable dcode_age has statistical significance for 3 regions

##	reg	scode	age	euler	dcode	dcode_age
## [1,]	0	0.96138159	2.752882e-10	0.0063404134	0.1294579	0.42360908
## [2,]	0	0.02596661	5.395799e-02	0.1395785747	0.7960038	0.03738268
## [3,]	0	0.05645537	0.000000e+00	0.0505620137	0.2452375	0.34434681
## [4,]	0	0.82496212	1.940831e-06	0.0002195698	0.6093196	0.87761639



Match-it dataframe (Age*Diagnosis) AFTER EXCLUDING DEVIANTS

```
p_val <- run_AgeDiagnosisModel(df_lC_matched,
                              measure = "CT_freesurfer",
                              Z = Zs_match,
                              exclude_deviants = T)

p_val_FDR <- Apply_FDR_Correction(p_val)
```

```
## Without FDR correction:
## Variable dcode has statistical significance for 28 regions
## Variable dcode_age has statistical significance for 12 regions
```

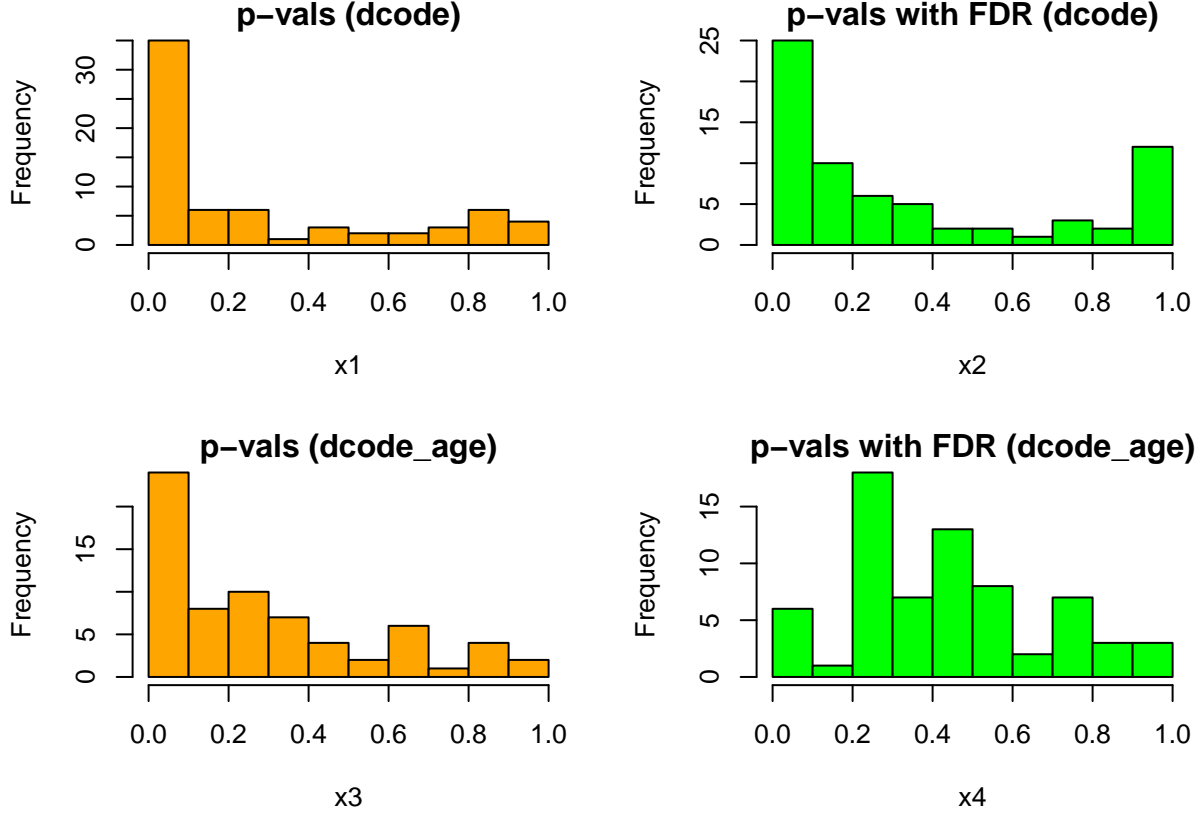
##	reg	scode	age	euler	dcode	dcode_age
## [1,]	0	0.74826138	2.994560e-11	7.108074e-06	0.04432257	0.1732372556
## [2,]	0	0.00674216	1.033507e-01	9.145237e-02	0.84604943	0.0007090065
## [3,]	0	0.00603199	0.000000e+00	7.779486e-03	0.12476858	0.4508306700
## [4,]	0	0.70418807	4.016180e-07	1.756316e-05	0.99929755	0.8471431413

```
## With FDR correction:
## Variable dcode has statistical significance for 20 regions
## Variable dcode_age has statistical significance for 5 regions
```

```

##      reg      scode      age      euler      dcode      dcode_age
## [1,]  0 0.83412744 6.170609e-11 4.394082e-05 0.1076405 0.39267111
## [2,]  0 0.03526669 1.115531e-01 1.515082e-01 0.9386440 0.01607081
## [3,]  0 0.03526669 0.000000e+00 1.726956e-02 0.2232701 0.57842426
## [4,]  0 0.80221857 6.068894e-07 8.855610e-05 0.9992976 0.88763773

```



Analysis of z-scores and computation of global scores

The following tables represent, **in terms of samples**, the number of deviations ($|Z| > 1.96$). It is calculated for each timepoint ($timepoint = 1, 2, 3$) and differentiated between controls and subjects.

NO MATCHED	timepoint 1	timepoint 2	timepoint 3
# total samples	31.688	31.348	10.744
# samples dev (%)	795 (2.51%)	696 (2.22%)	275 (2.56%)
# controls samples	20.264	19.924	7.412
# controls dev (%)	450 (2.22%)	335 (1.68%)	143 (1.93%)
# patients samples	11.424	11.424	3.332
# patients dev (%)	345 (3.02%)	361 (3.16%)	132 (3.96%)

MATCHED	timepoint 1	timepoint 2	timepoint 3
# total samples	22.916	22.304	6.664
# samples dev (%)	617 (2.69%)	519 (2.33%)	187 (2.81%)

MATCHED	timepoint 1	timepoint 2	timepoint 3
# controls samples	11.492	11.152	3.332
# controls dev (%)	269 (2.34%)	176 (1.58%)	71 (2.13%)
# patients samples	11.424	11.152	3.332
# patients dev (%)	348 (3.05%)	343 (3.08%)	116 (3.48%)

In the following, we chose to work with the matched dataset and the Zs derived from its lme model, providing a better statistical support for the analysis.

We computed the number of deviant samples ($Z < -1.96$ and $Z > 1.96$):

	Infra-normal deviants	Non deviants	Supra-normal deviants	Deviants
tp 1	339 (1.48%)	22.299 (97.31%)	278 (1.21%)	617 (2.69%)
tp 2	322 (1.44%)	21.785 (97.67%)	197 (0.88%)	519 (2.33%)
tp 3	133 (2.00%)	6.477 (97.19%)	54 (0.81%)	187 (2.81%)
total	794 (1.53%)	50.561 (97.45%)	529 (1.02%)	1.323 (2.55%)

We computed the number of subjects that are **deviant at timepoint 1** and stay **deviant in subsequent timepoints**, for each region. A maximum of 5 subjects (1.48% of total subjects) fulfill that condition for any region.

	Condition
rh_cuneus_CT_freesurfer	5
lh_rostralmiddlefrontal_CT_freesurfer	4
lh_temporalpole_CT_freesurfer	4

We computed the number of subjects that are **not deviant at timepoint 1** but become **deviant in subsequent timepoints**, for each region. A maximum of 20 subjects (5.93% of total subjects) fulfill that condition for any region.

	Condition
lh_lateraloccipital_CT_freesurfer	20
rh_temporalpole_CT_freesurfer	18
lh_posteriorcingulate_CT_freesurfer	18

1. Percentage of patients deviated from the normative range for any single cortical region

- Timepoint 1: No more than 6.548% of patients deviated from the normative range for any single cortical region.
- Timepoint 2: No more than 6.707% of patients deviated from the normative range for any single cortical region.

- Timepoint 3: No more than 10.204% of patients deviated from the normative range for any single cortical region.
-

2. Most common regions with infra-normal deviations. Percentage of patients.

- Timepoint 1: Infra-normal deviations in CT of subjects were most commonly located in **rh_fusiform_CT_freesurfer** cortices, although only 1.786% of patients showed significant deviations in these regions.
- Timepoint 2: Infra-normal deviations in CT of subjects were most commonly located in **rh_temporalpole_CT_freesurfer** cortices, although only 4.268% of patients showed significant deviations in these regions.
- Timepoint 3: Infra-normal deviations in CT of subjects were most commonly located in **lh_fusiform_CT_freesurfer** cortices, although only 8.163% of patients showed significant deviations in these regions.

DUDA: sacar la región y su porcentaje en los pacientes? Lo que hice fue sacar la región de los sujetos y el porcentaje para esa región concreta en los pacientes.

3. Most common regions with supra-normal deviations. Percentage of individuals.

- Timepoint 1: Supra-normal deviations in CT were most common in the **lh_lateraloccipital_CT_freesurfer** regions, 2.967% of individuals.
 - Timepoint 2: Supra-normal deviations in CT were most common in the **lh_lateraloccipital_CT_freesurfer** regions, 2.134% of individuals.
 - Timepoint 3: Supra-normal deviations in CT were most common in the **lh parahippocampal_CT_freesurfer** regions, 3.061% of individuals.
-

4. Percentage of subjects with at least one region infra-normal deviated. Patients vs Healthy controls.

- Timepoint 1: Infra-normal deviations for at least one region were evident in 29.167% of patients, whereas this was the case for 26.036% of healthy individuals.
 - Timepoint 2: Infra-normal deviations for at least one region were evident in 35.366% of patients, whereas this was the case for 27.439% of healthy individuals.
 - Timepoint 3: Infra-normal deviations for at least one region were evident in 40.816% of patients, whereas this was the case for only 38.776% of healthy individuals.
-

5. Percentage of subjects with at least one region supra-normal deviated. Patients vs Healthy controls.

- Timepoint 1: Supra-normal deviations for at least one region were evident in 32.143% of patients and 34.32% of healthy individuals.
 - Timepoint 2: Supra-normal deviations for at least one region were evident in 28.659% of patients and 29.878% of healthy individuals.
 - Timepoint 3: Supra-normal deviations for at least one region were evident in 30.612% of patients and 36.735% of healthy individuals.
-