Abstract

Individuals diagnosed with psychiatric disorders often have some degree of cognitive impairment as well. The relationship between psychiatric disorders and cognitive impairment is not straightforward, however. For instance, individuals with fewer years of education perform worse on cognitive tests and are also more likely to be diagnosed with a psychiatric disorder. In this study, we attempt to determine the extent to which diagnosis predicts a cognitive ability by taking into account each patient's demographic, fMRI, and MRI (structural images of the brain) data as well.

The dataset used in this study contains 220 observations of patients who have no psychiatric disorder (controls), as well as patients who have been diagnosed with ADHD, bipolar disorder, and schizophrenia. Multiple linear regression models were used to examine the relationship between the predictor variables and three response variables: verbal working memory (Verbal W.M.), spatial working memory (Spatial W.M.), and reaction time (Reaction). The model that used a combination of demographic, fMRI, and MRI data explained the most variability within the response variables, with R-squared values 0.323, 0.1884, 0.2041 for predicting Verbal W.M., Spatial W.M., and Reaction, respectively. Age (p-values < 0.01 for all three response variables) and Diagnosis-Control (p-values < 0.05 for all three response variables), as well as a number of MRI and fMRI variables were found to be statistically significant predictors of cognitive ability. Furthermore, models built on control patients showed a trend of over-predicting for diagnosed patients. Altogether the results indicate that there is a meaningful difference between diagnosed and control patients, which is partially explained by both diagnosis as well as demographic and structural factors.

Problems

In this project, the main question we want to answer is: *to what extent does a diagnosis predict a person's cognitive ability?* More specifically, there are three questions that can help us understand the relationship between diagnosis and cognitive ability:

1. Can we predict the relationship between cognitive ability using demographic, fMRI, and MRI data?
2. What can we learn by building a statistical model using only control subjects (i.e. subjects with no mental diseases) and using that model to predict cognitive ability on patients with mental health?
3. What other relationships can we find in the data set?

Variables of the Study

This study looks at total 220 observations (patients).

Demographic Variables

The following demographic information about the patients of the study were available for our analysis. Gender, language, race, and ethnicity were coded as categorical variables to better capture and understand the significance between different classes in each of these variables.

- Age

- Gender
- Language
- Main race
- Ethnicity
- School (years of)
- Diagnosis (DX)
    - Controls (no diagnosis of mental disorder)
    - ADHD
    - Schizophrenia
    - Bipolar Disorder

fMRI Variables

fMRI images show metabolic functions; these images are useful in analyzing connectivities and functions of different brain areas. The numerical fMRI scores available for this study are as follows:

- Visual global efficiency
- Somatomotor global efficiency
- Dorsal attention global efficiency
- Ventral attention global efficiency
- Limbil global efficiency
- Frontoparietal global efficiency
- Default mode global efficiency

Structural (MRI) Variables

MRI images show the anatomical structure of the brain. Numerical scores gathered from MRI images for this study are as follows:

- Left amygdala
- Right amygdala
- Left caudate
- Right caudate
- Left accumbens area
- Right accumbens area
- Total gray volume
- Cortex volume
- Cortical white matter volume
- Left putamen
- Right putamen
- Left pallidum
- RIght pallidum
- Left hippocampus
- Right hippocampus
- WM hypointensities
- Non-WM hypointencities

Response Variables (Cognitive Measures)

The following three scores were recorded for the patients as they performed a series of tasks. Higher verbal working memory and spatial working memory scores indicate that better performance, while lower reaction times indicate faster performance.

- Verbal working memory - the retention and manipulation of verbal information
- Spatial working memory - the retention and manipulation of visuospatial information
- Reaction time - time spent on a task

Exploratory Data Analysis

The exploratory data analysis includes three parts:

- Sample Structure
- Predictor Variable
- Response Variable

1. Sample Structures & Predictor Variables:

We observed that the CNP dataset we use divide individual sample into four categories including Control, SZ, ADHD and BD. So the first step of our exploratory data analysis is finding the distribution of samples among the four categories to see if there are problems such as unbalanced distribution. In general, we obtain a lot more samples in the control group that we do in the three other groups. Besides, different variables (we look into age, gender, race, ethnicity and language) all have their values distributed in an unbalanced way among the four categories.

Some of the unbalanced situations (like we got around 67% English speakers in the four group while only 33% of other languages) are really representative because they represent the realistic situation regarding the meaning of the predictor variable. While we still want to make further adjustments to decrease the effect of unbalanced distributions in other predictors.

2. Response Variables:

We also looked into the relationship between response variables and the four diagnosis groups. As in the second question, we need to study how diagnosis level would influence the model performance. After we plot the three response variables (Verbal.M, Spatial.M and Reaction time) with diagnosis types and set y-axi as distribution density value, we observed that control group generally has the highest density among different response variables. This kind of unbalanced situation is really similar to what we observed in predictor variables and we will talk about further observations and adjustments in the following parts.

Statistical Analysis

All the statistical analysis in this project was done by building multiple logistic regression models with the study data. The multiple logistic regression model was chosen since we had three numerical response variables, and a combination of categorical and numerical predictor variables.

Question 1:

First Model: Predictors: DX(diagnosis), Demographics.

Response Variables: Verbal Working memory, Spatial Working memory, Reaction.
Results from R:

1. DX and Demographics with Verbal Working Memory.

	Estimate	t-value	Pr(>|t|)		VIF
(Intercept)	0.141441	0.282	0.77851		
age	-0.034355	-4.794	3.09E-06	***	1.100384
gender	-0.030365	-0.234	0.81516		1.023988
language1	-0.084886	-0.625	0.5324		1.123344
race_main	0.051231	0.896	0.3715		1.516923
ethnicity	0.325741	1.891	0.05998	.	1.566826
DXBP	0.084317	0.394	0.69391		1.763061
DXControl	0.50656	2.76	0.00629	**	2.091299
DXSZ	-0.645778	-2.831	0.00509	**	1.85382

In this model, our significant variables are age and DX. The R-square of the model is 0.2748. From the VIFs, there is no collinearity between our predictors.

2. DX and Demographics with Reaction.

	Estimate	t-value	Pr(>|t|)		VIF
(Intercept)	-0.344044	-0.574	0.5669		
age	0.043825	5.12	6.88E-07	***	1.100384
gender	0.125757	0.812	0.4179		1.023988
language1	0.007683	0.047	0.9622		1.123344
race_main	-0.061355	-0.898	0.3702		1.516923
ethnicity	-0.372838	-1.812	0.0714	.	1.566826
DXBP	-0.11142	-0.436	0.6633		1.763061
DXControl	-0.531998	-2.427	0.0161	*	2.091299
DXSZ	0.169166	0.621	0.5353		1.85382

In this model, our significant variables are age and diagnosis. The R-square of the model is 0.1786. From the VIFs, there is no collinearity between our predictors.

3. DX and Demographics with Spatial Working Memory.

	Estimate	t-value	Pr(>|t|)		VIF
(Intercept)	-0.567375	-0.812	0.41795		
age	-0.028103	-2.817	0.00531	**	1.100384
gender	-0.047365	-0.262	0.79335		1.023988
language1	0.133593	0.707	0.48031		1.123344
race_main	0.156969	1.971	0.05	*	1.516923
ethnicity	0.038535	0.161	0.87247		1.566826
DXBP	0.215429	0.723	0.47026		1.763061
DXControl	0.831141	3.253	0.00133	**	2.091299
DXSZ	-0.758098	-2.388	0.01784	*	1.85382

In this model, our significant variables are age, race and diagnosis. The R-square of the model is 0.1983. From the VIFs, there is no collinearity between our predictors.

Second Model: predictors: fMRI, MRI, Demographics, DX.
Response Variables: Verbal Working memory, Spatial Working memory, Reaction.
Results from R:

1. fMRI, MRI, DX and Demographics with Verbal Working Memory.

	Estimate	t-value	Pr(>|t|)		VIF
(Intercept)	0.387658	0.747	0.45589		
age	-0.036521	-4.265	3.12E-05	***	1.676474
gender	0.010369	0.066	0.94716		1.585551
language1	-0.150385	-1.1	0.27284		1.217135
race_main	0.057129	0.984	0.32634		1.667832
ethnicity	0.239716	1.359	0.17587		1.755076
Left.Amyg	3.291396	1.781	0.07645	.	3.576404
Right.Am	-3.273935	-1.652	0.10023		4.006823
Left.Caud	1.441335	0.409	0.68305		12.200817
Right.Cau	-3.048329	-0.849	0.39712		13.101885
Left.Accu	-1.180882	-0.716	0.47475		2.853226
Right.Acc	3.151915	1.882	0.06133	.	2.976919
TotalGray	5.245809	0.869	0.38567		39.624271
CortexVol	-3.771683	-0.693	0.48886		31.746318
CorticalW	4.001783	2.71	0.00733	**	2.402985
Left.Puta	2.890851	1.218	0.22485		6.152861
Right.Put	-3.938545	-1.516	0.13114		7.355536
Left.Pallid	1.659782	1.071	0.28554		2.51303
Right.Palli	-3.784262	-2.286	0.02335	*	2.838914
Left.Hipp	-2.769714	-1.091	0.27647		6.776594
Right.Hip	0.675816	0.278	0.78132		6.165486
WM.hypo	-1.28505	-1.107	0.2695		1.345114
non.WM.	-1.187116	-1.154	0.24974		1.244913
DXBP	0.044046	0.205	0.83779		1.897768
DXControl	0.404545	2.102	0.03684	*	2.454507
DXSZ	-0.578388	-2.455	0.01498	*	2.111365

In this model, our significant variables are age, Cortical white matter volume, Right Pallidum and diagnosis. The R-squared of the model is 0.3205. From the VIFs, we detect collinearity between variables so we decided to drop two variables which have

similar biological functions towards cognitive abilities as the other two factors. The variables we dropped are Right caudate and Total gray volume.

2. fMRI, MRI, DX and Demographics with Reaction.

	Estimate	t-value	Pr(>|t|)		VIF
(Intercept)	-0.25344	-0.398	0.690772		
age	0.03528	3.361	9.36E-04	***	1.676474
gender	0.22748	1.187	0.236489		1.585551
language1	-0.05479	-0.327	0.744189		1.217135
race_main	-0.03378	-0.475	0.635606		1.667832
ethnicity	-0.40015	-1.85	0.06588	.	1.755076
Left.Amyg	1.63537	0.722	0.471267		3.576404
Right.Am	1.9575	0.805	0.421545		4.006823
Left.Caud	2.91795	0.675	0.500338		12.200817
Right.Cau	-1.77584	-0.403	0.687207		13.101885
Left.Accu	-0.28133	-0.139	0.889464		2.853226
Right.Acc	0.17081	0.083	0.933789		2.976919
TotalGray	10.83528	1.465	0.144597		39.624271
CortexVol	-13.80729	-2.071	0.039726	*	31.746318
CorticalW	-0.7901	-0.436	0.66304		2.402985
Left.Puta	-3.25899	-1.12	0.264264		6.152861
Right.Put	3.48798	1.095	0.274841		7.355536
Left.Pallid	-0.26066	-0.137	0.891033		2.51303
Right.Palli	-3.49848	-1.724	0.086376	.	2.838914
Left.Hipp	-1.87022	-0.601	0.548492		6.776594
Right.Hip	0.57308	0.192	0.847735		6.165486
WM.hypo	2.10375	1.479	0.140853		1.345114
non.WM.	-0.84092	-0.667	0.505553		1.244913
DXBP	-0.06669	-0.253	0.800395		1.897768
DXControl	-0.45314	-1.92	0.056271	.	2.454507
DXSZ	0.21775	0.754	0.451898		2.111365

In this model, our significant variables are age and cortex volume. The R-squared of the model is 0.1891. From the VIFs, we detect collinearity between variables so we decided to drop two variables which have similar biological functions towards cognitive abilities as the other two factors. The variables we dropped are Right caudate and Total gray volume.

3. fMRI, MRI, DX and Demographics with Spatial Working Memory.

	Estimate	t-value	Pr(>|t|)		VIF
(Intercept)	-0.32597	-0.437	0.66252		
age	-0.03345	-2.718	7.17E-03	**	1.676474
gender	0.02445	0.109	0.9134		1.585551
language1	0.05389	0.274	0.78425		1.217135
race_main	0.18108	2.17	0.03122	*	1.667832
ethnicity	-0.13512	-0.533	0.5948		1.755076
Left.Amyg	6.14065	2.312	0.02183	*	3.576404
Right.Am	-5.03097	-1.766	0.079	.	4.006823
Left.Caud	-0.75066	-0.148	0.88236		12.200817
Right.Cau	-2.02771	-0.393	0.69492		13.101885
Left.Accu	-2.08848	-0.881	0.37927		2.853226
Right.Acc	6.01525	2.499	0.01329	*	2.976919
TotalGray	7.37959	0.851	0.39582		39.624271
CortexVol	-6.56983	-0.84	0.40172		31.746318
CorticalW	2.26516	1.067	0.28721		2.402985
Left.Puta	-2.82154	-0.827	0.40934		6.152861
Right.Put	-1.5345	-0.411	0.68156		7.355536
Left.Pallid	0.79705	0.358	0.72088		2.51303
Right.Palli	-2.63386	-1.107	0.26971		2.838914
Left.Hipp	-5.13648	-1.408	0.16068		6.776594
Right.Hip	3.80872	1.09	0.27707		6.165486
WM.hypo	-1.43041	-0.858	0.39216		1.345114
non.WM.	0.80047	0.542	0.58871		1.244913
DXBP	0.2492	0.807	0.42068		1.897768
DXControl	0.95289	3.445	0.0007	***	2.454507
DXSZ	-0.50282	-1.485	0.13923		2.111365

In this model, our significant variables are age, race, left amygdala, right accumben area and diagnosis. The R-squared of the model is 0.1992. From the VIFs, we detect collinearity between variables so we decided to drop two variables which have similar biological functions towards cognitive abilities as the other two factors. The variables we dropped are Right caudate and Total gray volume.

4. fMRI, MRI, DX and Demographics with Verbal Working Memory after dropping variables.

	Estimate	t-value	Pr(>|t|)		VIF
(Intercept)	0.404279	0.781	0.43578		
age	-0.035878	-4.214	3.83E-05	***	1.664043
gender	0.003304	0.021	0.9831		1.582771
language1	-0.166671	-1.228	0.22101		1.20369
race_main	0.05143	0.897	0.37075		1.632098
ethnicity	0.247852	1.411	0.15991		1.746354
Left.Amyg	3.253786	1.772	0.07791	.	3.543326
Right.Am	-3.490179	-1.773	0.07777	.	3.96557
Left.Caud	-1.03049	-0.723	0.47037		2.000726
Left.Accu	-1.46072	-0.909	0.36436		2.718688
Right.Acc	3.225316	1.951	0.0525	.	2.911706
CortexVol	0.699467	0.425	0.67158		2.922517
CorticalW	4.064538	2.763	0.00627	**	2.392771
Left.Puta	3.492176	1.518	0.13075		5.801754
Right.Put	-3.857685	-1.494	0.13667		7.288487
Left.Pallid	1.589207	1.028	0.30516		2.508733
Right.Palli	-3.918562	-2.441	0.01553	*	2.678838
Left.Hipp	-2.445102	-1.01	0.31368		6.187506
Right.Hip	0.855431	0.354	0.72359		6.107409
WM.hypo	-1.434055	-1.246	0.21409		1.327006
non.WM.	-1.277077	-1.253	0.21174		1.227718
DXBP	0.035497	0.166	0.86852		1.892134
DXControl	0.411197	2.172	0.03106	*	2.383943
DXSZ	-0.582235	-2.51	0.01287	*	2.053194

In this model, our significant variables are age, Cortical white matter volume, Right Pallidum and diagnosis which are the same significant variables as our previous model. The R-squared of the model is 0.323. From the VIFs, we detect no collinearity between variables.

5. fMRI, MRI, DX and Demographics with Reaction after dropping variables.

	Estimate	t-value	Pr(>|t|)		VIF
(Intercept)	-0.23532	-0.37	0.71184		
age	0.0366	3.498	5.80E-04	***	1.664043
gender	0.21992	1.148	0.25216		1.582771
language1	-0.07627	-0.457	0.64804		1.20369
race_main	-0.03476	-0.493	0.62227		1.632098
ethnicity	-0.39912	-1.849	0.06601	.	1.746354
Left.Amyg	1.75521	0.778	0.43752		3.543326
Right.Am	1.60438	0.663	0.50794		3.96557
Left.Caud	1.9096	1.091	0.27674		2.000726
Left.Accu	-0.40459	-0.205	0.83784		2.718688
Right.Acc	0.03598	0.018	0.98589		2.911706
CortexVol	-4.51123	-2.229	0.02697	*	2.922517
CorticalW	-0.78258	-0.433	0.66552		2.392771
Left.Puta	-2.64602	-0.936	0.35058		5.801754
Right.Put	3.82865	1.207	0.22889		7.288487
Left.Pallid	-0.37045	-0.195	0.84557		2.508733
Right.Palli	-3.31322	-1.68	0.09463	.	2.678838
Left.Hipp	-0.75878	-0.255	0.79892		6.187506
Right.Hip	0.64401	0.217	0.82844		6.107409
WM.hypo	1.93178	1.366	0.17338		1.327006
non.WM.	-1.05089	-0.839	0.40249		1.227718
DXBP	-0.08693	-0.33	0.74148		1.892134
DXControl	-0.47552	-2.044	0.04229	*	2.383943
DXSZ	0.17454	0.612	0.54097		2.053194

In this model, our significant variables are age and cortex volume and diagnosis. The R-squared of the model is 0.1884. From the VIFs, we detect no collinearity between variables.

6. fMRI, MRI, DX and Demographics with Spatial Working Memory after dropping variables.

	Estimate	t-value	Pr(>|t|)		VIF
(Intercept)	-0.3107	-0.418	0.676332		
age	-0.03254	-2.662	8.40E-03	**	1.664043
gender	0.01803	0.081	0.935831		1.582771
language1	0.03706	0.19	0.849364		1.20369
race_main	0.17846	2.169	0.031317	*	1.632098
ethnicity	-0.13156	-0.522	0.602505		1.746354
Left.Amyg	6.18651	2.347	0.019905	*	3.543326
Right.Am	-5.28841	-1.872	0.062757	.	3.96557
Left.Caud	-2.17904	-1.065	0.287988		2.000726
Left.Accu	-2.25474	-0.978	0.329445		2.718688
Right.Acc	5.97527	2.518	0.012612	*	2.911706
CortexVol	-0.24957	-0.106	0.916051		2.922517
CorticalW	2.29238	1.086	0.278959		2.392771
Left.Puta	-2.2902	-0.693	0.488953		5.801754
Right.Put	-1.33392	-0.36	0.719248		7.288487
Left.Pallid	0.71576	0.323	0.747357		2.508733
Right.Palli	-2.59142	-1.125	0.262144		2.678838
Left.Hipp	-4.45932	-1.283	0.200883		6.187506
Right.Hip	3.91135	1.128	0.260642		6.107409
WM.hypo	-1.57211	-0.952	0.342316		1.327006
non.WM.	0.66184	0.452	0.651535		1.227718
DXBP	0.23588	0.767	0.443808		1.892134
DXControl	0.94418	3.474	0.000631	***	2.383943
DXSZ	-0.52587	-1.58	0.11583		2.053194

In this model, our significant variables are age, race, left amygdala, right accumben area and diagnosis which are the same significant variables for previous model without modification. The R-squared of the model is 0.2041. From the VIFs, we detect no collinearity between variables.

Question 2:

For problem two, we built two linear models using only controls (patients with no mental disorders) to predict the cognitive ability of patients with mental disorders. The first model consists of only the demographic data. The second model contains the demographic data as well as MRI and fMRI data, excluding factors with high multicollinearity we inspected in Question1.

Question 3:

For question 3, we used the PCA method to inspect the main factors that caused the variations in the MRI and fMRI data of patients with different mental disorders and healthy patients. Currently, we have 24 variables for MRI and fMRI data when we only have 220 observations. However, these variable could be linearly dependent and highly correlated. Using the PCA method, we can reduce the dimension of these 24 variables and extract the most important factors while preserving the highest variations in the data.

We also conducted a machine learning method, KNN, to predict the mental disorders of patients based on their fMRI, MRI and cognitive ability data. We wanted to see whether

we can determine what mental disorder a patient has based on his or her cognitive ability data, MRI and fMRI results.

Summary of the Results

Question 1: Can we predict the relationship between cognitive ability using demographic, fMRI, and MRI data?

Among the nine models, it is obvious that the only significant variable that exists in all models. There are some other significant factors as well, however according to their p-value, they are not as strong as the factor age.

Comparing the R-squared table that is attached below, we can see that When using data from all patients, model with Demographics, DX, fMRI, MRI predictors after dropping the highly correlated variables give the best fit with adjusted R^2 = 0.323. Moreover, after dropping the two highly correlated variables, two of the adjusted R^2 increased, which indicates that our models after modification provide a better fit.

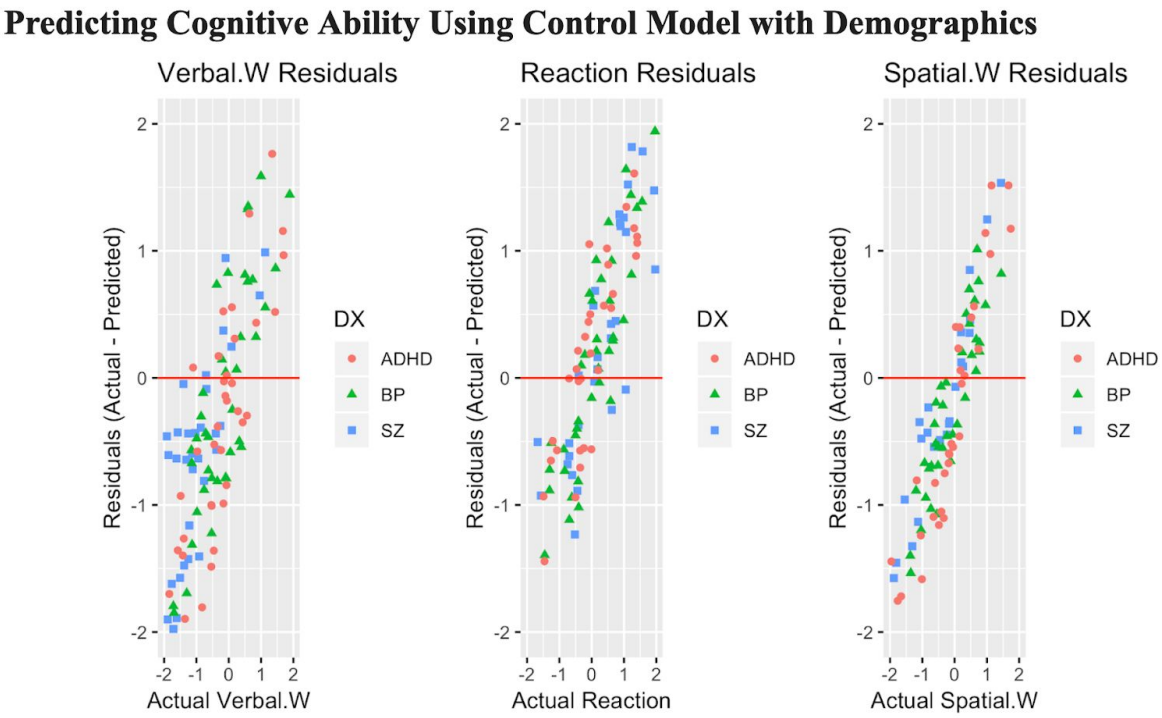
	Demographic + DX	Demographic + DX + fMRI + MRI Before	Demographic + DX+ fMRI + MRI After
Verbal W.M.	**0.2748**	**0.3205**	**0.323**
Reaction	**0.1786**	**0.1891**	**0.1884**
Spatial W.M.	**0.1983**	**0.1992**	**0.2041**

Question 2:

For the first model containing only demographics variables, for the verbal working memory, we have an adjusted R squared of 0.1403, which is considered reasonable for this type of data. For the other two cognitive ability measures, our adjusted R squared is not too optimal and is yet to be improved.

	Adjusted R^2
Verbal W.M.	**0.1403**
Reaction	**0.1085**
Spatial W.M.	**0.04679**

Additionally, we found trends where the control model was consistently over-predicting response variables for the diagnosed patients. This pattern can be seen by plotting the residuals (actual - predicted) over actual scores:



Below are tables indicating the number of observations where the model over-predicted (i.e. where the residual between the actual and predicted results is negative):

i.e., are the residuals (actual - predicted) < 0?

Verbal.W

	ADHD	BP	SZ
No	12	6	6
Yes	**25**	**26**	**32**

Reaction

	ADHD	BP	SZ
No	**24**	**26**	**25**
Yes	13	16	13

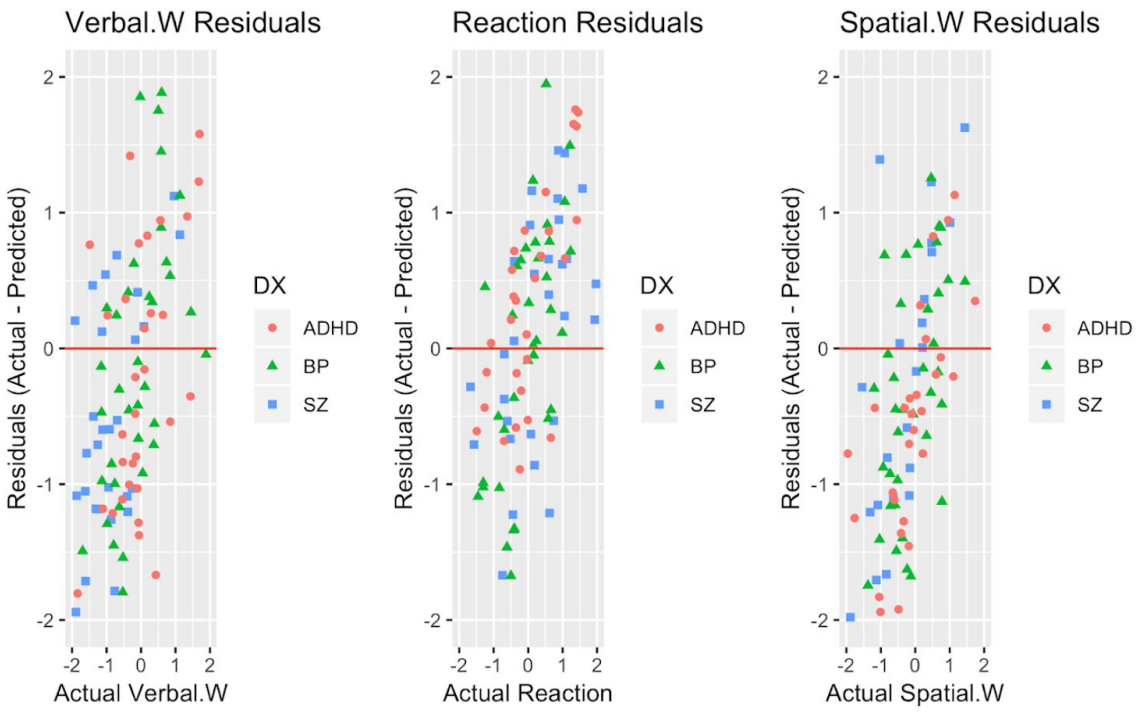
Spatial.W

	ADHD	BP	SZ
No	13	14	9
Yes	**24**	**28**	**29**

Similarly for the second model with the demographic, MRI and fMRI variables, we have a reasonable adjusted R squared for the verbal working memory while having slightly lower adjusted R squared for the reaction time and spatial working memory. Moreover, for the working memory, we found the most significant and influential variables to be age and ethnicity. For reaction time, we found that cortex volume has the most significant impact, while for spatial working memory, we found the most significant variables to be race and left amygdala.

	Adjusted R^2
Verbal W.M.	**0.1524**
Reaction	**0.08428**
Spatial W.M.	**0.1136**

Similar to the first model, the second control model also over-predicted the response variables for the diagnosed patients.



Below are the number of observations the control model over-predicted:

Verbal.W

	ADHD	BP	SZ
No	14	16	10
Yes	**23**	**26**	**28**

Reaction

	ADHD	BP	SZ
No	**25**	**27**	**25**
Yes	12	15	13

Spatial.W

	ADHD	BP	SZ
No	9	13	11
Yes	**28**	**29**	**27**

Question 3:

From our PCA results, we can determine that 3 to 5 principal components can most accurately represent the variations of the data. These factors include total gray volume, right putamen, frontopolar and ventral_a. These factors contributed to the greatest variations among patients with no mental disorder and patients with bipolar disorder, ADHD and schizophrenia.

For our KNN predictions, we achieved 53% accuracy. Although the accuracy rate is not yet optimal, considering the dimensions and unbalance of our data, the result is reasonable. We can improve the accuracy of this algorithm by reducing our dimensions by taking only the most important factors that we found through the PCA.

Interpretation of the Results

Question 1:

Four highly correlated variables were found: total gray volume, total cortex volume, left caudate and right caudate. Conceptually, total cortex volume includes the gray volume. Left caudate and right caudate are symmetrical in the brain and serve a similar function in the memory processing. Those can explain their high VIF. So we excluded Total Gray Volume and Right Caudate from all the models.

Our models have given credit to those significant variables, their corresponding functions to individual cognitive measurements can be found in literatures. Age is found universally significant. Studies have shown that older adults put more effort into focusing during encoding, in order to compensate for a reduced ability to hold information in working memory.Cortex Volume is found significant in models with reaction time as the outcome variable. Cortex volume can be seen as an indicator of the size of brain connectivities. Intuitively, bigger brain means smarter brain. Left Amygdala is mainly associated with spatial working memory. Amygdala is the emotion center. As memory is stored with emotions, it serves an important role in memory processing

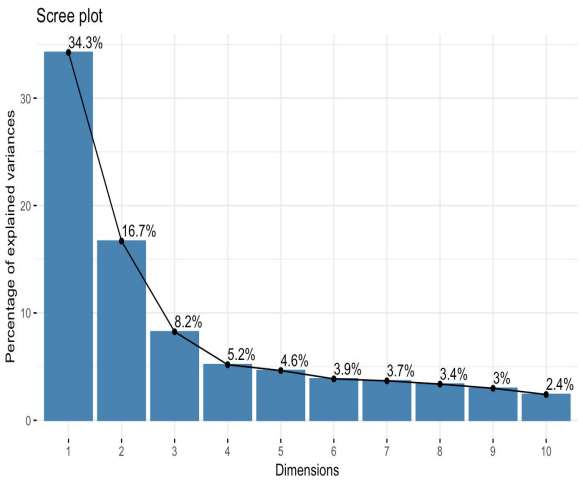
Question 2:

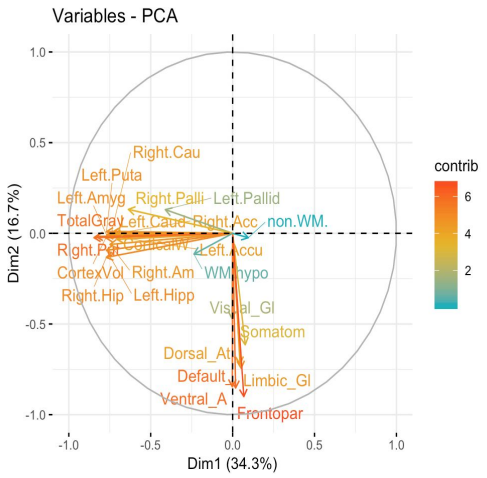
The pattern of residuals between the control model's predicted scores and the diagnosed patients' actual scores indicate that there is some meaningful difference between the control and diagnosed scores. The model build on control patients expected diagnosed patients to have *higher* Verbal.W and Spatial.W scores than they actually had, and *faster* Reaction times than they actually had. Based on the results from Question 1, it seems that some part of this difference *is* due to the diagnosis of the patients, since diagnosis was a statistically significant predictor for cognitive ability in many of the models for Question 1.

Interpretation of Plots

Question 3:

The first plot we have is the principal component analysis plot. From this plot, we can see that around three to five factors contributed most significantly to the variations of the MRI and fMRI data among patients with different mental disorders. Therefore, we can pick the top three to five variables and extract them from the pool of the 24 MRI and fMRI variables.





From the second plot, we can see that the variables that contributed most significantly to the variations among the patients are total gray volume, right putamen, frontopolar and ventral_a.

Overall Conclusions

Age is the only significant variable that exists in all models. Model with Demographics, DX, fMRI, MRI predictors after dropping the highly correlated variables give the best fit with adjusted R^2 = 0.323. The control model consistently over-predict response variables for the diagnosed patients.

Challenges of the Study

Initially coming in to this experiment, we expected that we can find some strong key factors from fMRI and MRI that affect cognitive abilities such as verbal working memory, spatial working memory and reaction since we are provided with research questions associated with those variables. Therefore, considered the number of variables and their biological relationships, we use multiple linear regression for our models. Later on, for the models we built upon, the only strong predictor that is significant among all models, is age. The result is reasonable, since according to Caroline N. Harada, Marissa C. Natelson Love and Kristen Triebeld's study, "Cognitive change as a normal process of aging has been well documented in the scientific literature. Some cognitive abilities, such as vocabulary, are resilient to brain aging and may even improve with age. Other abilities, such as conceptual reasoning, memory, and processing speed, decline gradually over time" (Harada). However, our goal of finding key factors from fMRI and MRI variables is not achieved. To our understanding, the main reason why we couldn't reach a crucial conclusion for our first research question is due to our data. The data is mostly unbalanced with more females than males, more hispanic than non-hispanic and a relatively small age range. Even though our models have a high R-squared in social science studies, the R-square is not high enough for real life.

Recommendations for the Future

For our future improvements, we would like to build models with consideration of interaction effects between all kinds of demographic factors such as sex, age, etc. Also, we would like to collect more external data to balance our data, in order to reach a more precise results and a better fit of our model (higher R-squared). If possible, we would like to try more regression methods and prediction methods to compare.

Citations

Harada, Caroline N, et al. “Normal Cognitive Aging.” *Clinics in Geriatric Medicine*, U.S. National Library of Medicine, Nov. 2013, www.ncbi.nlm.nih.gov/pmc/articles/PMC4015335/.