Methods

Data on 243 people is provided. The data contains information on age, Symbol Digit Modalities Test (SDMT) score, and CAG repeat length for healthy controls and people with pre-manifest Huntington's Disease.

Differences between the variance in SDMT across the two groups are tested using an F test. Differences in mean SDMT score across the two groups are tested using an independent two-sample t-test with equal variances. The equal variance assumption is justified from the results of the F test, see results section.

Pearson correlations between SDMT and age are calculated in each group. Age adjusted comparisons between the groups are made using ANCOVA. All continuous variables are centered before being used in ANCOVA models to aid interpretability.

Several linear models are fit on the pre-manifest HD group. Associations between age, CAG repeat length, and SDMT score are calculated and used to explain differences in the coefficient estimates from the models.

Finally, SDMT scores are compared across three groups. The groups are – healthy controls (group 1), people with CAG repeat length up to and including 42 (group 2), and people with CAG repeat length greater than 42 (group 3). ANOVA is used to compare mean SDMT score across the three groups, and ANCOVA is used to adjust for the effect of age in these comparisons.

Assumptions for all models were checked by examining the distribution of residuals. This was done graphically by plotting residuals against each variable in the model, and by looking at QQ-plots.

Results

Question 1

There are 120 pre-manifest Huntington's Disease gene carriers and 123 healthy controls in the data. Table 1 shows summary statistics for these two groups.

	Age (years)		**SDMT**		**CAG repeat length**
	Healthy Control	Pre-manifest HD	Healthy Control	Pre-manifest HD	Pre-manifest HD
Range	23 - 66	19 - 64	30 - 78	30 - 80	39 - 52
Mean	46.2	40.8	52.2	51.3	43.1
Std Dev	10.3	8.9	9.5	10.2	2.4
Skew	0	0.3	-0.1	0.2	1
Kurtosis	2.2	3	2.9	2.7	4.5

Table 1. Summary statistics

HD – Huntington's Disease, SDMT – symbol digit modalities test score, Std Dev – Standard Deviation
Higher values of SDMT are better, maximum value possible is 110
CAG repeat length not recorded in controls

Age & Symbol Digit Modalities Test (SDMT) score are approximately normally distributed in both groups. Healthy controls have a higher mean age compared to the pre-manifest group, and the age distribution in healthy controls shows some evidence of light tails.
CAG repeat length is right skewed and has heavy tails, with a kurtosis of 4.5. The high kurtosis in CAG repeat length is mostly due to three observations (id 133, 151, and 167).

There is no clinical reason to class these observations as outliers however, so they were not removed in any of the following analyses.

Question 2

A two-sided F-test shown no significant difference between the standard deviation of SDMT scores between the two groups at the 95% level (F = 0.86 on 122 & 119 degrees of freedom, p = 0.41). As there is no evidence of significant differences in variance between the groups, a two-sided, independent, equal variance t-test was used to assess differences in mean SDMT score across the groups. The t-test shown no significant difference between mean SDMT scores in the healthy control & pre-manifest HD groups at the 95% level (t = 0.73 on 241 degrees of freedom, p = 0.46).

There is a negative correlation between age & SDMT scores. In healthy controls the correlation is -0.39 (95% CI [-0.53, -0.23]), and in the pre-manifest HD group it is -0.13 (95% CI [-0.30, 0.06]). There is no significant difference between the correlation across the groups, but they both show similar effects – as age increases, SDMT scores tend to decrease. As the healthy control group is older than the pre-manifest HD group, it is possible that the mean SDMT score in the healthy control group is biased downwards and is lower than it should be. This bias will reduce the difference in means.
Table 2 shows the means & differences. Adjusting for age using an ANCOVA model increases the difference from 0.9 to 2.4:

	Healthy Control	Pre-manifest HD	Difference
Unadjusted	52.2	51.3	0.9 (1.3)
Adjusted for age	52.9	50.5	2.4 (1.3)

Table 2. Mean SDMT scores by group

Standard errors for the difference estimate shown in brackets

Residual checks show that the assumptions in the ANCOVA model are met. There is some evidence that the constant variance assumption is violated, with group 2 having a larger residual variance compared to group 1. A QQ-plot shows that the residuals aren't exactly normal for particularly large negative & positive residuals. Both observations are evidence that the assumptions are violated, but the violations are not so large to make results from the model invalid.

Question 3

3 models were fit to examine the relationship between age, CAG repeat length, and SDMT score in the pre-manifest HD group. Coefficient estimates are shown in Table 3:

	Model 1	Model 2	Model 3
Constant	51.26 (0.93)	51.26 (0.94)	51.26 (0.92)
Age	-0.15 (0.11)		-0.46 (0.18)
CAG repeat length		-0.03 (0.39)	-1.42 (0.67)

Table 3. Coefficient estimates from linear models

Standard errors shown in brackets
All models fit using data on the pre-manifest HD group only
Age and CAG variables have been centered

Model 3 shows that age and CAG repeat length both have a significant negative effect on mean SDMT score – for example holding all else constant, increasing age by one year will decrease mean SDMT score by 0.46 on average. These effects weren't seen in models 1 or

2 because there is a confounding relationship between age and CAG repeat length.
By just looking at age alone, model 1 compares mean SDMT scores between people of the same age but with potentially very different CAG repeat lengths. Similarly model 2 compares mean SDMT scores between people with similar CAG repeat lengths but potentially very different ages.
There is a strong negative correlation (-0.82, 95% CI [-0.87, -075]) between age and CAG repeat length. There are negative correlations between age and SDMT score (-0.13, 95% CI [-0.30, 0.06]) and CAG repeat length and SDMT score (-0.01, 95% CI [-0.19, 0.17]) also. This explains why the coefficient estimates in models 1 & 2 are lower than in model 3. Take model 1 for example – it correctly identifies that SDMT score decreases as age increases, but people with high age values in group 2 tend to have lower CAG repeat lengths. Given the negative correlation between CAG repeat length and SDMT score, we would expect these lower CAG repeat length people to get higher SDMT scores. The effect of CAG repeat length is 'washing out' some of the age effect due to the correlations. A similar explanation also holds for model 2.

Looking at the residuals against CAG repeat length, all three models show that the constant variance assumption does not hold. There is also evidence that the larger residuals do not follow a normal distribution from QQ-plots. This will affect the robustness of results in Table 3 – in particular, the estimates for standard errors may not be accurate.

Question 4

Table 4 shows some summary statistics across three groups (defined in the methods section, and repeated in the table):

	Group 1 (N = 123) Healthy Controls			**Group 2 (N = 49)** CAG ≤ 42			**Group 3 (N = 71)** CAG > 42		
	Age	CAG	SDMT	Age	CAG	SDMT	Age	CAG	SDMT
Range	23 – 66		30 – 78	34 – 64	39 – 42	30 – 72	19 – 50	43 – 52	31 – 80
Mean	46.2		52.2	48.2	41.0	51.1	35.7	44.6	51.4
Std Dev	10.3		9.5	7.1	0.9	9.8	5.9	2.0	10.6
Skew	-0.1		-0.1	0.3	-0.4	-0.2	-0.4	1.7	0.5
Kurtosis	2.2		2.9	2.6	2.1	2.3	3.2	5.7	2.9

Table 4. Summary statistics

CAG – CAG repeat length, SDMT – symbol digit modalities test score, Std Dev – Standard Deviation
CAG repeat length not recorded in healthy controls
Group size (N) shown at top of table columns

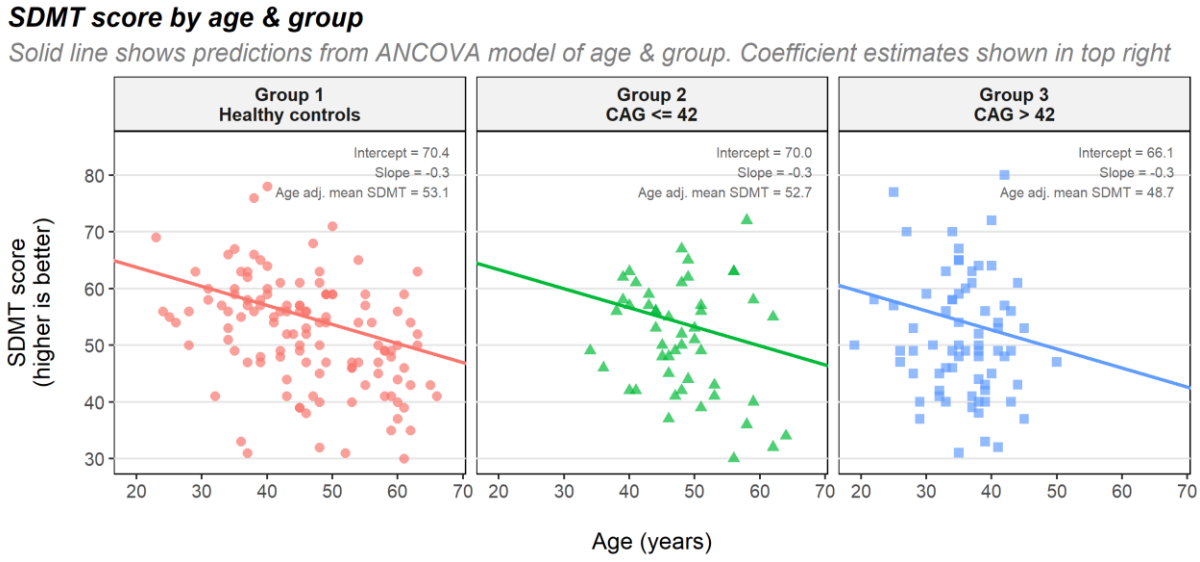
The distribution of SDMT scores are similar across the three groups. Splitting CAG repeat length at 42 has roughly split the pre-manifest HD group into a younger (CAG > 42) and older group. The distribution of CAG repeat length differs in groups 2 and 3. Group 3 has a much larger spread of CAG repeat lengths compared to group 2.

A one-way ANOVA found no significant difference between the mean SDMT scores in the three groups.
As we saw in the previous section, age and CAG repeat length are confounders and need to be adjusted for in any comparisons across groups. Adjusting for age using ANCOVA shows that there is a significant difference in SDMT – the mean SDMT score in healthy controls is significantly higher compared to group 3. After adjusting for age, the mean SDMT score in group 3 is 4.3 units lower than group 1 (95% CI for the difference: [-7.5, -1.2]).

Residual plots show no large violations of the assumptions for these models. The QQ-plot for the ANCOVA model shows that the residuals more closely follow a normal distribution compared to the ANOVA model, suggesting that the results from the ANCOVA model are

more robust. There is some evidence that the constant variance assumption does not hold in both models, but this is very minor with almost no pattern in the residuals.

The chart shows the predictions from the ANCOVA model, along with the actual values. Model estimates for each group, including the age adjusted SDMT scores, are also shown:



Discussion

This report examined the relationship between SDMT score, age, and CAG repeat length. CAG repeat length is only recorded in the pre-manifest HD group. There are negative correlations between age and SDMT score, between CAG repeat length and SDMT score, and between age and CAG repeat length. These correlations mean that age is a confounder and should be adjusted for when comparing SDMT scores across groups. Adjusting for age shown that health controls have a higher mean SDMT score compared to the pre-manifest HD group. Further analysis shown that people with a high value of CAG repeat length (above 42 in our analysis) have a significantly lower mean SDMT score compared to health controls. This suggests that the symbol digits modalities test may be a useful clinical tool in identifying Huntington's disease. The test may be useful for screening, for example by developing a reference range of typical scores for each year of age. If a person who is suspected to be at risk of Huntington's disease scores below this range, then the clinician may decide to send the patient for follow up genetic tests to confirm or rule-out the diagnosis. Whether this screening tool has any clinical benefit would depend on many factors – such as the costs of developing the tool, the potential benefits in earlier diagnosis, and the potential harms of false positives (e.g. emotional stress from a positive screening test) and false negatives (e.g. delays in treatment caused by delayed diagnosis). Exploring these topics is not possible with the current data.

The models considered in this report could be improved. The final model shown some evidence that the constant variance assumption did not hold. This could potentially be improved by collecting additional data on other variables, such as sex or familial risk factors. More complex models may be considered as well. For example, Table 3 shown that there are associations between age, CAG repeat length, and SDMT score, so a model which included interactions between age & CAG repeat length may have been insightful. The final model categorised CAG repeat length, from a continuous variable into a discrete variable. Other models such as splines may have better captured the relationships explored in this model.

Appendix – model formula

The model in (4c) is

$$\mathrm{y_i} = \alpha + \beta_1 G_i + \beta_2 \mathrm{H_i} + \beta_3 \mathrm{age}_i + \epsilon_i$$

Where

- y_i= Predicted SDMT score for the person i in the data
- α = Mean SDMT score in group 1 (healthy controls)
- β_1 = Difference in mean SDMT score between healthy controls and group 2 (people with CAG repeat length ≤ 42)
- G_i = Indicator variable, 1 if person i is in group 2 and 0 otherwise
- β_2 = Difference in mean SDMT score between healthy controls and group 3 (people with CAG repeat length > 42)
- H_i = Indicator variable, 1 if person i is in group 3 and 0 otherwise
- β_3 = The expected change in mean SDMT score for a 1 unit increase in age, holding all other variables constant
- age_i = The age of person i in the data
- ϵ_i = Error term, assumed to be normally distributed with mean zero and constant variance

Age was centered before fitting the model, so all the means mentioned above are when age is equal to its mean value (43.5).