Dementia Risk Factors: EDA of easySHARE data

Michelle Cleary, Fionnuala Marshall, Ellen Crombie

Executive Summary

Year on year, the number of people with dementia is rising, along with the associated global dementia cost. The purpose of this report is to investigate modifiable risk factors of dementia and determine whether dementia prevention policies may be effective. Using data from high income countries (HICs), namely Belgium and Austria, we evaluated and validated factors suggested by the 2017 Lancet Commission [1].

The results of our statistical experiments agreed with wider research literature, that older age has a very significant impact on the onset of dementia. In light of this, we aimed to determine whether modifiable risk factors such as education level reached in early life, drinking behaviour and quality of life index should be modified by participants to reduce the risk of dementia.

The results suggest that the further the education level reached in early life (younger than 45 years), the lower the risk of dementia in later life. The biggest improvement seen in reducing dementia severity is from primary to secondary education, as seen in Figure 1. Although this conclusion is drawn from HIC data where primary and secondary education is almost always compulsory, it may in fact be more relevant to low- and middle-income countries (LMICs), where secondary education is not necessarily provided or mandatory. In this case, policy should prioritise ensuring secondary education as a mandatory or incentivised option for all.

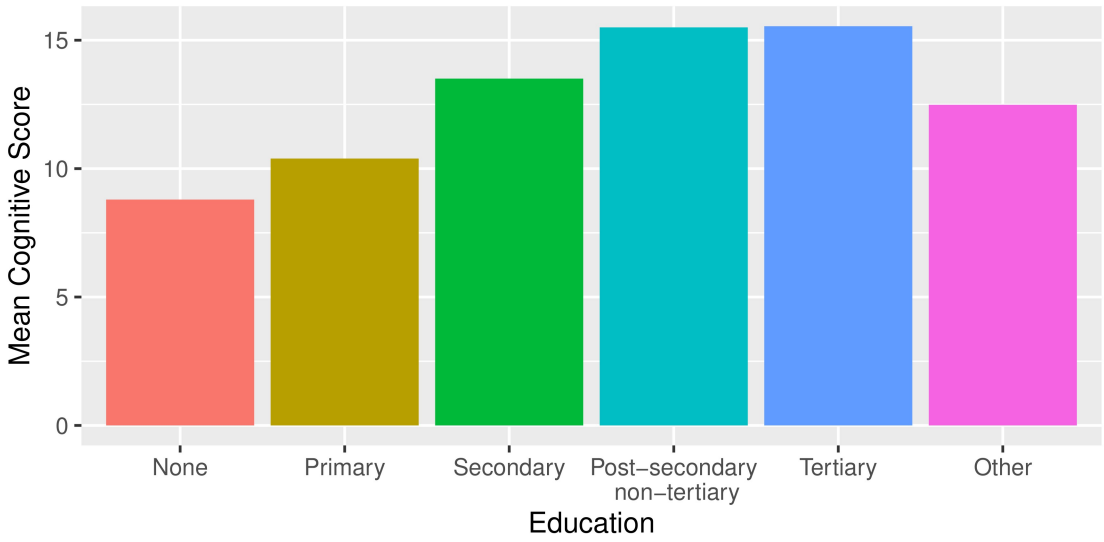

Figure 1: Education level against mean cognitive score, where cognitive score is used as a proxy for severity of dementia. A lower cognitive score indicates higher severity of dementia. The figure shows that lower levels of education indicate a lower cognitive score.

Further to this, difficulty concentrating is associated with an increased risk of the onset of dementia in later

life, as shown in Figure 2. Therefore keeping the brain active throughout life would be beneficial, whether this be through adult education, playing board games such as chess, or joining a book club.

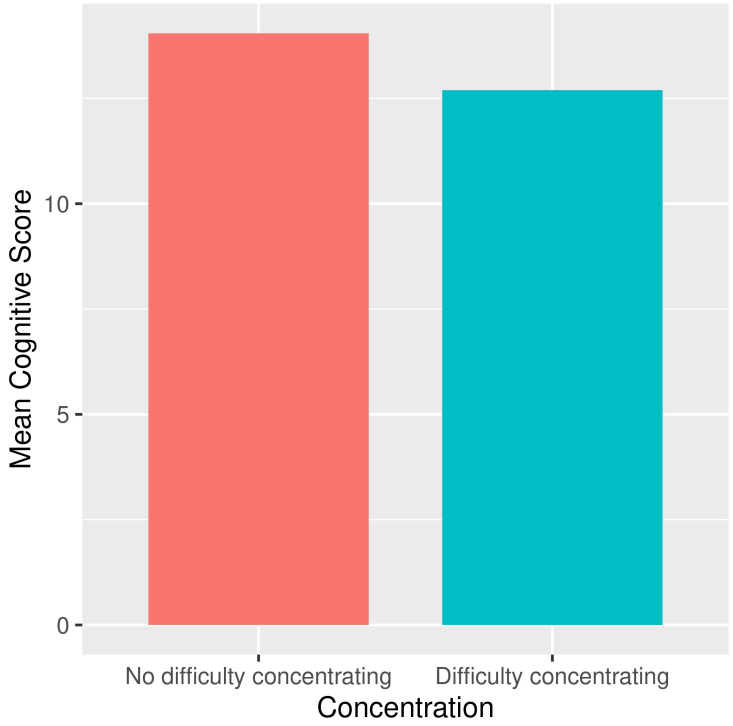

Figure 2: Concentration level against mean cognitive score, where cognitive score is used as a proxy for severity of dementia. A lower cognitive score indicates higher severity of dementia. The figure shows that difficulty concentrating indicates a lower cognitive score.

This report ensures the model was validated across data from different years and across different countries. However the main focus is on data from HICs, where health policies are developed and care is well established on the whole. It would be of particular use to investigate whether the suggested preventative measures are indeed relevant to LMICs. Populations in these countries are reaching later ages more regularly than in previous decades, and so preventative policies may have large impact on reducing the risk of dementia.

Key messages

1. The number of people with dementia is rising globally, however the majority of research investigates data from High Income Countries.
2. Older age has a significant association with a higher risk of dementia but there are modifiable risk factors which might prevent or delay dementia.
3. Recommended preventative policies include providing sufficient primary and secondary education for all children as this may reduce the risk of dementia in later life.
4. Keeping the brain active throughout life may also reduce risk of dementia. There are many ways to achieve this, including adult education, card games, and reading.

Introduction

It has been estimated that almost 9.9 million people develop dementia each year, and by 2030 it is predicted that the current 55 million people affected will have risen to 75 million [2]. The WHO report entitled 'The Epidemiology and Impact of Dementia' [3] describes vast and extensive research into older age as a risk factor of dementia, however the Lancet Report [1] identifies 12 modifiable risk factors using evidence from high-income countries (HICs). These include; social isolation, highest education level reached and excessive alcohol consumption.

In this report, we will create a model for dementia risk prediction using the easySHARE data set, and assess the predictive performance of our model using validation. We will focus on a selection of modifiable risk factors of dementia as indicated by the Lancet Report, and aim to determine whether these risk factors as a whole might outperform a benchmark model which uses only age as the predictor.

Karel G M Moons' review article on Risk Prediction Models [4] confirms that validation is particularly important, and warns that good performance of a selected model on a selected set of data is insufficient, even after internal validation. This report aims to remedy the inadequacies of common medical research by performing spatial and temporal validation on the model, using clear and consistent data handling.

We will analyse and validate data from HICs, namely Belgium and Austria. Therefore, preventative measures and risk-reducing suggestions may differ in LMIC cultures and settings.

Individual contributions include:

- Michelle Cleary (s1979093): Assessment of predictive performance and fitting the model.
- Ellen Crombie (s1907212): Executive summary and pre-processing.
- Fionnuala Marshall (s1907509): Model assumption validity and graphical summaries.

Methods

We investigated the risk factors of dementia using data from Belgium within Wave 4 of the easySHARE data set. This wave was selected based on the low rate of missing information and high participation levels within relevant questionnaire questions. The initial risk factors were selected by their relevance to the 12 modifiable predictors determined by the Lancet Report [1]; education, hearing loss, traumatic brain injury, hypertension, alcohol consumption, smoking, obesity, physical activity, depression, social isolation, diabetes and air pollution.

Creating the response variable

EasySHARE does not record diagnosis of Alzheimer's disease (dementia) in all waves. Instead, we created a composite cognitive score as a proxy for dementia severity, ranging from 0 (bad) to 25 (good), combining the recall scores and numeracy measures, following Crimmins' Assessment of cognition study [5]. It is worthwhile to note that we take the mean value of both numeracy scores from the data set to make the model applicable to further waves where just one or both of the numeracy measures were available. The assessment of cognition study [5] suggests using orientation as a component of cognitive score, however, this report chooses not to do so due to the vast amount of missing values within the data set (11,406 out of the possible 20,370 for waves 4 and 5 at this stage of analysis).

Cleaning and selecting the data

We extracted 28 variables to investigate from the 102 available, including gender, in order to explore whether certain risk factors differed for male and female participants. As suggested by the risk predicition article [4], certain variables were omitted if more than 10% of entries were missing. Figure 3 shows that, from our chosen subset of variables, only 1.5% of the remaining observations have missing values. This implies that the model

we create using these variables will have a higher level of integrity and be a more accurate prediction from the variables selected. For example, the number of years of education completed was not included because 1575 out of the possible 20370 entries were missing. Instead, the furthest level of education reached was included, which had fewer missing values (246) and could describe the same effect.

Further variables were not selected on this basis. These included the residential proximity of children and the number of grandchildren and siblings alive, which may have been important contributors in determining a respondent's social interaction. However, in a high income country such as Belgium, they may have had less of an effect due to improved ease of transport.

Next, we coded factor variables and made simplifications based on the number of entries at each level. For example, furthest education levels were combined into 6 levels from the original 9. The level 'still in school' was dropped due to it including only 6 data values, all of which were from respondents who were aged 50 and below and outside the main focus area of this investigation. Factor levels for living area were also reduced to distinguish between living in an urban or rural area, with 12492 and 6990 entries for each level respectively and 888 missing values.

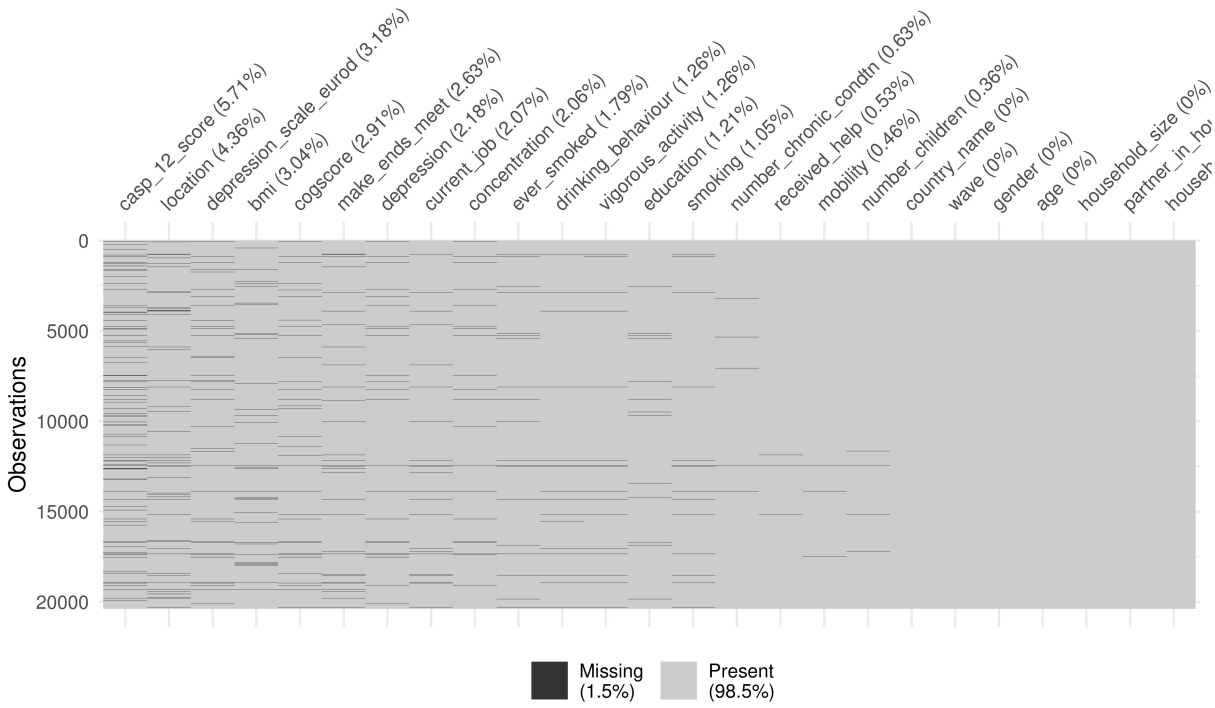

Figure 3: Plot sorting variables by proportion of missing values

Fitting the model

We separated 200 randomly selected observations (the validation data), with the remaining 5122 making up the training data set. When then fit a linear model based on the training data.

The significance level ($p < 0.01$) was chosen to reduce the risk of selecting less important predictors. However, as detailed by the Risk Predicition article [4], predictor inclusion and exclusion was not solely dependent on the statistical significance of each predictor. Instead, a combined approach of evaluating statistical significance and performing stepwise selection was employed. Both AIC and BIC criterion were trialled, yet we eventually chose to use BIC criterion since the sample size was sufficiently large and a consistent, simpler model was favourable [6].

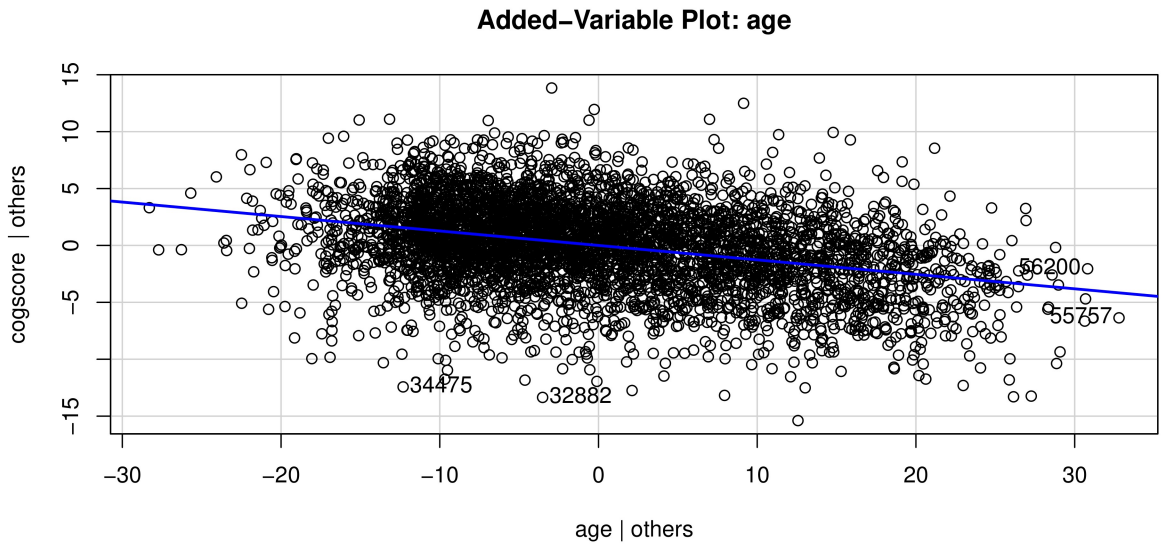

Figure 4: Scatterplot of age against cognitive score, holding all other predictors constant

Figure 4 confirms the findings of the WHO Epidemiology report [3] and clearly identifies an association between increased age and a lower cognitive score. Although at each age there is a wide spread of cognitive score values, the average scores show a strong negative correlation with age.

Results

Final model

We model cognitive score, cogscore_i, for individual $i = 1, ..., n$ as

$$\text{cogscore}_i = \alpha + \boldsymbol{x}_i^T \boldsymbol{\beta} + \epsilon_i$$

with ϵ_i following a normal distribution with mean zero and variance σ^2. The vector $\boldsymbol{x}_i$ is a row of the model matrix for observation i containing dummy variables for the factor variables gender, education, concentration, and drinking_behaviour. In addition, $\boldsymbol{x}_i$ contains the continuous variable age and the discrete variable casp_12_score. As identifiability constraints, the first level of each factor is set to zero. The vector $\boldsymbol{\beta}$ contains all the parameters relating to the columns in the model matrix X. That is, $\boldsymbol{\beta}$ is given by the estimate values in Table 1, which illustrates the summary statistics for the final model. Table 2 provides explanations of factor levels and meaning of variable names.

Table 1: Regression coefficient estimates, 95% confidence intervals, and p-values of the predictors in the final model for cognitive score

	Estimate	2.5%	97.5%	p-value
(Intercept)	16.6306	15.2094	18.0517	0.0000
gender1: female	0.5987	0.3855	0.8118	0.0000
age	-0.1270	-0.1373	-0.1168	0.0000
education1 - Primary education	1.0695	-0.0025	2.1414	0.0505
education2 - Secondary education	3.0088	1.9532	4.0645	0.0000
education3 - Post-secondary non-tertiary education	4.6964	1.8812	7.5117	0.0011
education4 - Tertiary education	4.6719	3.6063	5.7375	0.0000
education5 - other	2.8788	1.2415	4.5162	0.0006
casp_12_score	0.0432	0.0256	0.0608	0.0000
concentration1: difficulty concentrating	-0.7914	-1.0525	-0.5302	0.0000
drinking_behaviour2 - Once or twice a week	0.6180	0.3369	0.8990	0.0000
drinking_behaviour3 - Three or four days a week	0.6987	0.3032	1.0942	0.0005
drinking_behaviour4 - Almost every day	0.6231	0.3611	0.8851	0.0000

Table 2: Explanatory variables

Variables	Explanation	Type	Levels
gender	Gender	Factor, 2 levels	Male, female
age	Age	Continuous	-
education	Level of education	Factor, 6 levels	None, primary, secondary, post-secondary non-tertiary, tertiary, other
casp_12_score	CASP-12 score (quality of life), ranging from 12 to 48	Discrete	-
concentration	Concentration	Factor, 2 levels	No difficulty with concentration, difficulty with concentration
drinking_behaviour	Drinking behaviour	Factor, 4 levels	Less than twice a month or almost never, once or twice a week, three or four days a week, almost every day

All of the predictor variables within our model are statistically significant. Being female is associated with a 0.5987 increase in cognitive score compared to being male. CASP-12 score has a weak positive association with cognitive score, with a 0.0432 increase in cognitive score per unit CASP-12 score increase. Surprisingly, more frequent drinking behaviour have positive associations with cognitive score. Higher levels of education also suggest a higher cognitive score. Age and concentration are associated with cognitive score decreasing by 0.1270 and 0.7914 per unit increase respectively.

Validation of assumptions

As seen in Figure 5, we plotted a graph of Residuals vs Fitted Values in order to check if our model assumption of constant variance was valid.

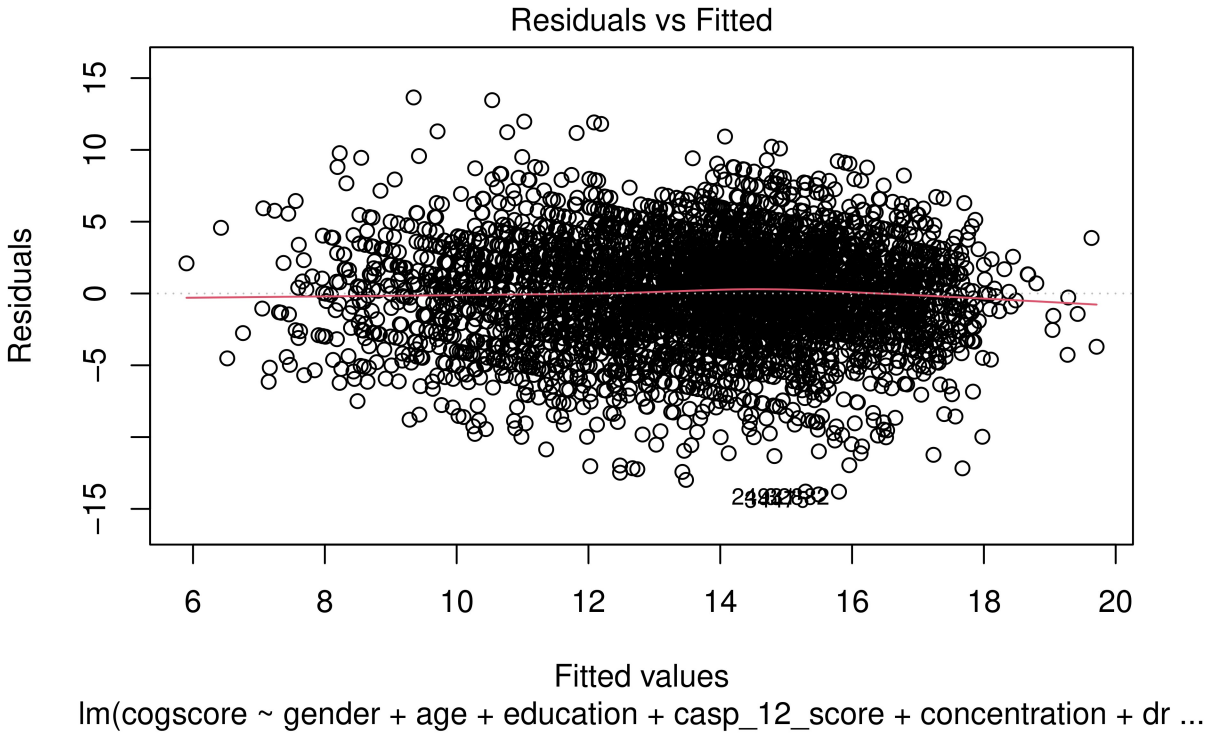

Figure 5: Residuals vs Fitted values plot to test model assumption of constant variance

The residuals appear to be randomly scattered around the zero line. This implies that the zero expectation of residuals for our model appears to be entirely justified.

When investigating the normality assumption of our model, we found there were slight deviations in the tails of the Q-Q plot. Upon adjusting the model by transforming variables, this made no significant difference to either plot. Thus, creating a more complex model to very insignificantly bring the model closer to our required assumptions did not appear to be a sensible decision.

Assessment of predictive performance

We assess the predictive performance of the model by estimating root mean squared error (RMSE), which compares the true and predicted cognitive score of each individual. We also estimate the mean Dawid-Sebastiani (MDS) score, which compares each true value to the mean and variance of the predicted values for cognitive score. Both of these scores are negatively oriented, i.e. the lower the score, the better.

Letting n be the total number of observed individuals in the test dataset and y_i the true predicted cognitive score of individual i, we define the scores as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2},$$

where $\hat{y}_i$ is the predicted cognitive score of individual i, and

$$MDS = \frac{1}{n}\sum_{i=1}^{n}\frac{(y_i-\mu)^2}{\sigma^2} + \log(\sigma^2),$$

where μ and σ^2 are the prediction mean and variance, respectively.

We use the following test data for validation:

- Training data - Belgium, Wave 4.
- The validation data (200 observations left out from initial data) - Belgium, Wave 4.
- The training data, but a different wave - Belgium, Wave 5.
- The training data, but a different country - Austria, Wave 4.

We define a benchmark model as:

$$\text{cogscore}_i = \gamma + \omega(\text{age}_i) + \epsilon_i$$

with ϵ_i following a normal distribution with mean zero and variance σ^2, γ = 24.0394, and ω = -0.1595.

We compare the scores for our model with those for the benchmark model to assess our model's predictive performance and the importance of variables for prediction.

Table 3: Root mean squared error for each test dataset

	Benchmark model	Actual model
Training data: Belgium - Wave 4	3.9472	3.5375
Validation data: Belgium - Wave 4	4.0728	3.4156
Belgium - Wave 5	3.9619	3.5771
Austria - Wave 4	4.4422	3.8169

As seen in Table 3, the RMSE for our actual chosen model was lower than that of the benchmark model for each dataset, indicating that the variables in our model are important for prediction. Focusing on our chosen model, all of the RMSE values are quite similar, with a difference of 0.4013 between the highest and lowest. The RMSE was lowest for the validation data. Particularly, it is lower than that for the training data, which suggests we have not overfitted the model. The RMSE for the test data from Wave 5 is very similar to that for the training data, indicating that our model's predictive performance is similar across different waves. Although the test data for Austria had the highest RMSE, it is still relatively low. This suggests that our model's predictive performance may not be as accurate across different countries, but is still relatively good.

Table 4: Mean Dawid-Sebastiani score for each test dataset

	Benchmark model	Actual model
Training data: Belgium - Wave 4	4.2530	3.9773
Validation data: Belgium - Wave 4	4.4866	3.9579
Belgium - Wave 5	4.3192	3.9899
Austria - Wave 4	5.0089	4.1044

Similar to RMSE, the MDS for our model was lower than that of the benchmark model for each dataset, shown in Table 4. This strengthens our previous observation that the variables in our model are important for prediction.

With regard to our chosen model, all of the MDS values are quite similar, with the test data for Austria having the highest score, and the validation data having the lowest. The scores again suggest that overfitting is not an issue, and that our model's predictive performance is similar across different waves and relatively good across different countries. Overall, the MDS scores provide further support to the conclusions drawn from the RMSE values above.

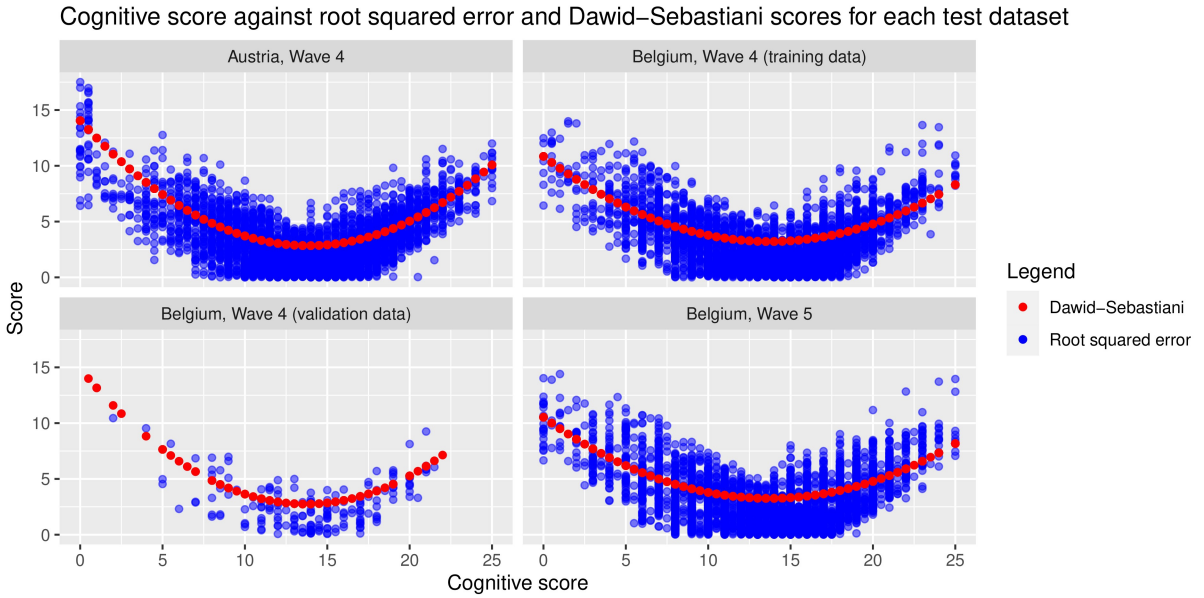

Figure 6: Scatterplot of cognitive score against root squared error and Dawid-Sebastiani scores for each test dataset, illustrating that both the root squared error (RSE) and Dawid-Sebastiani (DS) scores follow the same U-shaped trend.

Figure 6 shows that both root squared error and Dawid-Sebastiani scores are highest for extreme value cognitive scores. They are lowest for mid-range cognitive scores, between 10-15. This suggests that our model is best at predicting mid-range cognitive scores, and is less accurate at predicting the more extreme valued scores. These results are not surprising as only 2.2% and 3.9% of the observations in the training data had cognitive score values below 5 and above 20, respectively. Since our sample had very few extreme value scores and many mid-range scores, we would expect it to be most accurate at predicting mid-range scores.

Conclusion

The model highlighted several key findings. Firstly, as can be seen in Figure 7, difficulty in concentration is associated with a negative cognitive score and therefore a higher severity of dementia, as opposed to not having any difficulty concentrating. Whilst improving concentration is not a straightforward task, we believe this is a modifiable risk factor that can be alleviated by keeping cognitviely active in later life. This may include adult education or learning, playing thinking games, or joining a book club.

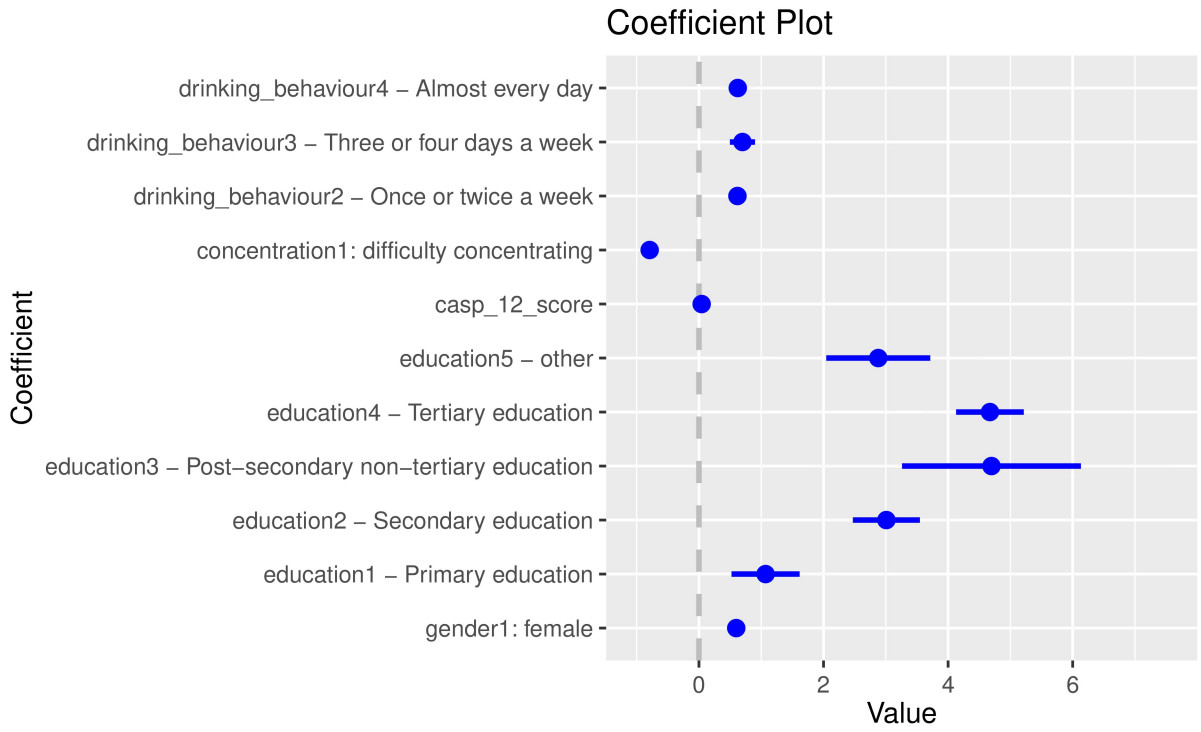

Figure 7: Plot of coefficients and standard errors for factor variables from the final model

The further the level of education reached in early life is also associated with a higher cognitive score in later life, and therefore a lower severity of dementia. Figure 8 shows that mean cognitive score continually increases from primary through to tertiary levels of furthest education reached. Considering this, we recommend that preventative policies for dementia should indeed prioritise primary and secondary education, and perhaps investigate ways to encourage tertiary education where possible.

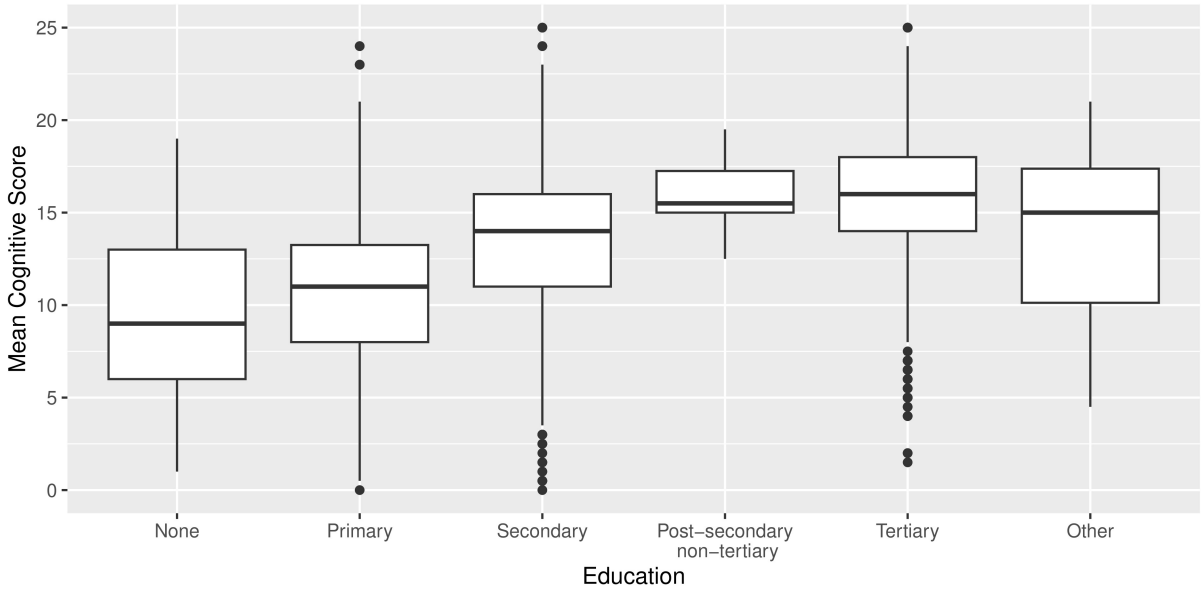

Figure 8: Education level against cognitive score boxplot, showing that higher levels of education are associated with a higher cognitive score.

When modelling with particular variables, we came across some variables which were poorly measured. This consequently made modelling with them cause some difficulties as it was not clear how different subgroups were defined. A clear example of this was within the education classification variable. Here, a category was labelled 'other', however, upon looking at all remaining categories, it was not clear what other levels of education may have not already been included. This factor level of the education variable was significant within our model, nonetheless, without extra information, it was impossible to draw conclusions and explanations regarding what this might mean in relation to cognitive score.

Whilst consistent data handling and statistical methods have been applied throughout the formation of this model, there are limitations that should be addressed. In particular, it would have been beneficial to validate the model over a larger time frame, in addition to the temporal validation already carried out (between Waves 4 and 5). Modifiable risk factors such as concentration and drinking behaviour were assessed during the same period of each participant's life. However, data regarding these behaviours from when the participants were younger (below 45 years old) may have been relevant in explaining the onset of dementia.

For example, this model suggests that frequent drinking behaviour has a positive association with cognitive score compared to less than twice a month or not at all, as can be seen in Figure 7 and visually in Figure 9. On the other hand, the Lancet Commission indicates increased frequency of drinking as a likely risk factor of dementia. One possible reason for this discrepancy is that the data set provides no information about the participants' drinking behaviour throughout their early life. Drinking frequently at a young age may or may not increase the risk of dementia, even if participants drink with a reduced frequency in later life. This is something our model cannot currently determine.

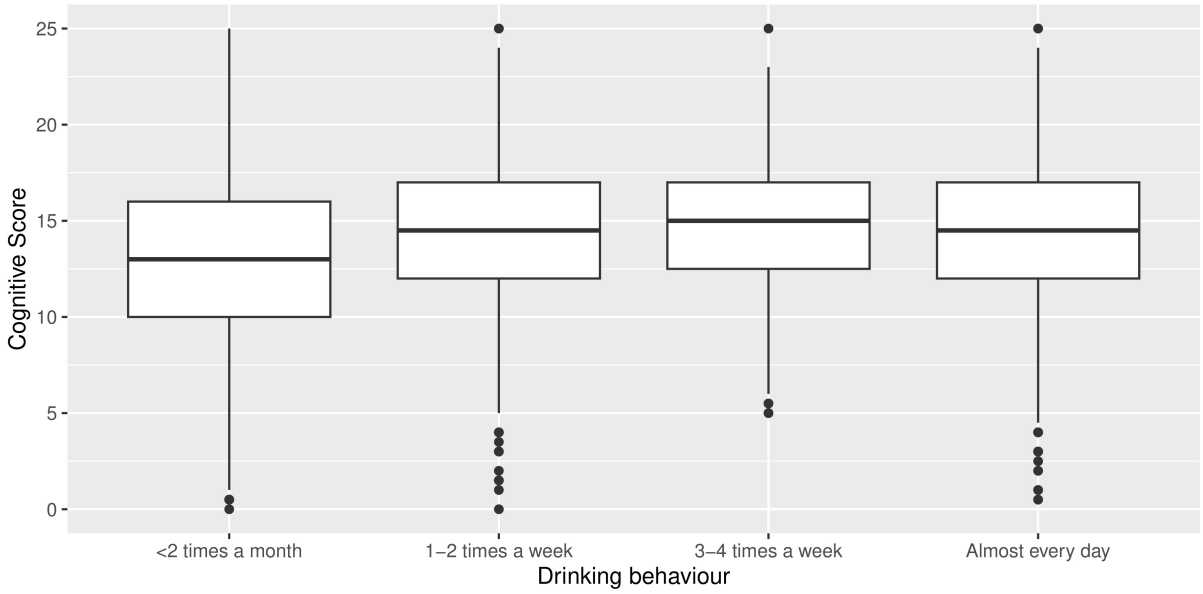

Figure 9: Drinking behaviour against cognitive score boxplot, showing that increasing frequency of drinking is associated with a higher cognitive score.

In conclusion, the model has confirmed that older age is a significant predictor of dementia, and has identified further modifiable risk factors such as difficulty in concentration and lower levels of education reached. Furthermore, this report has assessed the predictive performance of the model using validation to ensure good performance.

References

[1] Livingston, Gill et al., *Dementia prevention, intervention, and care: 2020 report of the Lancet Commission*, The Lancet, Volume 396, Issue 10248, 413 - 446, 2020

[2] World Health Organization, *Global action plan on the public health response to dementia 2017–2025*, https://www.who.int/publications/i/item/9789241513487

[3] Prince et al., *The Epidemiology and Impact of Dementia - Current State and Future Trends. WHO Thematic Briefing*, https://www.researchgate.net/publication/277217355_The_Epidemiology_and_Impact_of_Dementia_-_Current_State_and_Future_Trends_WHO_Thematic_Briefing, March, 2015

[4] Moons KGM, Kengne AP, Grobbee DE, et al., *Risk prediction models: II. External validation, model updating, and impact assessment*, https://heart.bmj.com/content/98/9/691, Volume 98, 2012

[5] Crimmins et al., *Assessment of cognition using surveys and neuropsychological assessment: the Health and Retirement Study and the Aging, Demographics, and Memory Study*, https://pubmed.ncbi.nlm.nih.gov/21743047/, 2011

[6] Henry de-Graft Acquah, *Comparison of Akaike information criterion (AIC) and Bayesian information criterion (BIC) in selection of an asymmetric price relationship*, https://academicjournals.org/article/article1379662949_Acquah.pdf, December, 2009