Assignment 3, Part 1

- *Code and data can be found here* JanCoUnchained
- Branch:Master

Model selection:

We chose to predict acoustic features by the fixed effects diagnosis and study. Study is included as a fixed effect (as opposed to a random effect) because we need to make sure that the studies do not differ significantly. We also included random slopes for ID and triangle. We included an interaction between diagnosis and study to make sure that the change in features due to diagnosis across studies was not significant:

```
lmer(acoustic_feature ~ diagnosis * study + (1|ID) + (1|triangle), data = data)
```

Diagnosis significantly predicted mean and median. We expect mean and median f0 to be very correlated parameters, so it is unsurprising that they would be predicted similarly.

Effect of study:

We also found a significant main effect of study 2 on mean, median, sd, iqr, mean_abs. Study 4 significantly predicted the parameter of range. This is problematic since it indicates that the same experimental paradigm yielded significantly different results on different occasions. We would have preferred the different studies to not differ significantly. It is somewhat misleading to include study as a fixed effect, since we are not expecting (nor do we desire) systematic variance to be caused by this predictor. However, we included it to test whether there were significant differences. It is at this moment uncertain whether study 2 is the issue, or whether it is study 1 and study 3. There are many potential explanations of the difference, one of them being that the conductor of the second experiment might have had a lower mean f0 and induced a mimicking effect in the participants.

Interaction effect: There are interaction effects between study and diagnosis for virtually all the acoustic parameters. For dependent variables mean, standard deviation, range and median there is an interaction between study 1 and 4.

Interpretation of the predictors

It is curious that study 1 & 2 differ as main predictors, while study 1 & 4 interact with diagnosis. This indicates that in study 2 the dependent variables are similarly affected in both groups in a way that is significantly different from in study 1. In study 4 one group is affected significantly more on dependent variables compared to study 1.

Model Results

Models	Dependent Variable	Significant fixed effects	Significant Interaction
Model 1	Mean	Diagnosis 1, Study 2	Diagnosis 1 & Study 4
Model 2	Standard Deviation	Study 2	Diagnosis 1 & Study 4
Model 3	IQR	Study 2	Diagnosis 1 & Study 2
Model 4	Range	Study 4	Diagnosis 1 & Study 4
Model 5	Median	Diagnosis 1 & Study 2	Diagnosis 1 & Study 4
Model 6	Mean Absolute Deviation	Study 2	-

Conclusion:

We are left with results that are quite difficult to interpret; *Study* seems to be a significant effect, albeit an unwanted one. Without further knowledge of how the individual studies were conducted, we cannot safely say what has caused this significance. This significance sadly intervenes with our interpretations of other significant effects, making further conclusions seem invalidated.

Assignment 3, Part 2

- *Code and data can be found here* JanCoUnchained
- Branch:Master

Task 1

Data Extraction

Data was extracted during the last assignment.

Summary and additional info: Last time we gathered descriptive statistics about the vocal pitch of schizophrenics and controls.

We obtained information about the range, median, mean, coefficient of variation and interquartile range.

The dataset from last time was combined with the provided clinical data, which ultimately resulted in the removal of some participants (n = 6). 2 of which had no clinical data, 4 which had no pitch data.

Integrating the clinical data in our analysis allowed us to include more random effects, such as gender, in our models of pitch range.

A new variable: "uniqueness" was added to distinguish between matched participants, as participants shared ID.

Comparison of performance

We used cross-validation to compare performance of our models. Refer back to Assignment 1, Part 3 for a conceptual overview of this method. Specifically, for this assignment we conducted our cross-validation in the following way:

We identified the relevant metrics for our different models. Then, we partitioned the data set into folds.

The partitions were split in a manner that grouped participants based on their ID variable. In other words, matched participants were grouped into the same folds, which also ensured that our partitions consisted of a balanced ratio of schizophrenics and controls.

Each model was then fitted five times, each time using data in the fold as the training set and the rest of the partitioned data as the testing set. Each iteration of each model was then used to predict diagnosis from the testing set.

We computed performance measures for the model predictions. Brier scores were calculated for each model, which allowed us to select the model with the best performance.

Task 2, Model Design

Range as a predictor

In order to test whether including random effects in the model bettered it, we made three models: one without random effects, one with study as a random intercept, and one with ID and study as random intercepts.

We can see from the cross-validation that the model without random effects outperforms both models with random intercepts on all performance measures. The model with both ID and study as random intercepts are particularly bad.

As can be seen on the visualised confusion matrix, this model has basically no true negatives and very few true positives. Cross-validations yields slightly better performance measures except for AUC than simply using a single train/test-split.

Looking at the different measures, the results are not very impressive. An accuracy of 57.4% is far too low to be of much value.

The NPV is a bit better than the PPV indicating that cases where people who are labelled non-schizophrenic are actually non-schizophrenic are more prevalent than cases where people who are schizophrenic are actually schizophrenic.

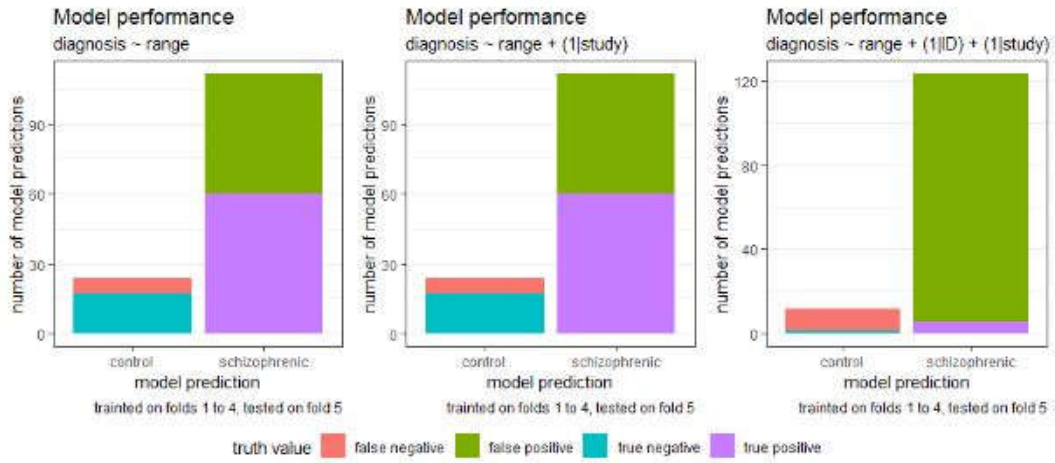
There is also a large discrepancy between the sensitivity and specificity of our predictive model which means that we correctly classify 72.3% of schizophrenic people, but we also mislabel many non-schizophrenic people.

If we we care less about mislabelling non-schizophrenic people and our goal is simply to "catch" schizophrenics, this balancing is generally positive. But the measures are still too low for our model to be of much to be of much use.

Figure 11: Models

model	SEN	SPE	PPV	NPV	AUC	ACC
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1 diagnosis ~ range	0.723	0.426	0.562	0.623	0.575	0.574
2 diagnosis ~ range + (1|study)	0.666	0.409	0.527	0.581	0.537	0.536
3 diagnosis ~ range + (1|ID) + (1|study)	0.238	0.074	0.198	0.074	0.844	0.155

Figure 12: Model Performance with range as a predictor



Best single predictor

Below is a table of the best performing model on various metrics:

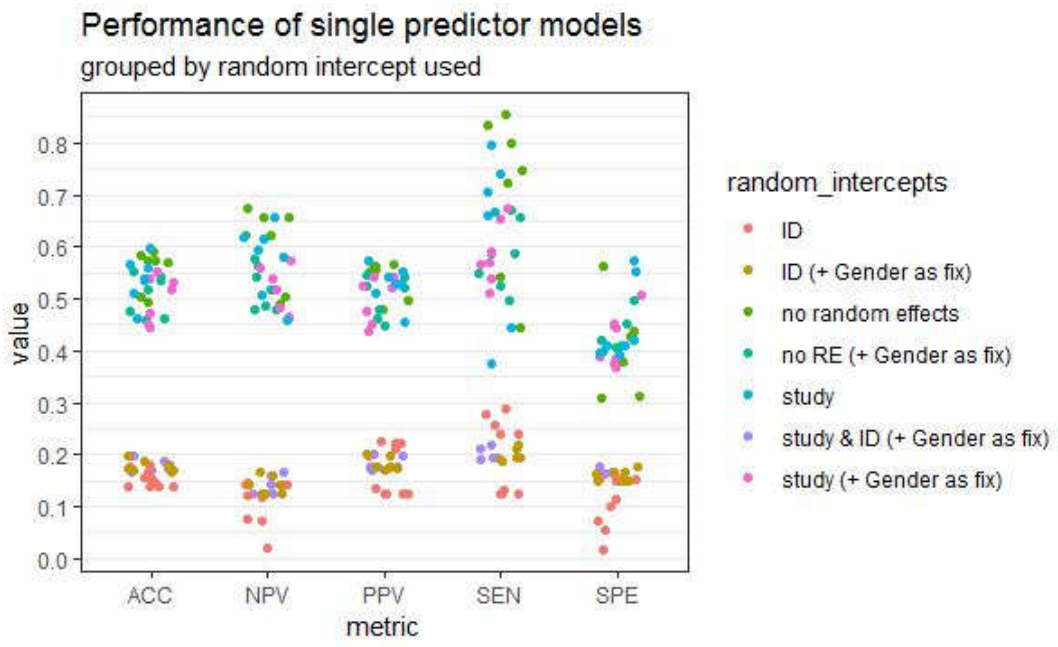
Measure	Model
Highest sensitivity	diagnosis ~ IQR
Highest sensitivity	diagnosis ~ mean + (1/study)
Highest PPV	diagnosis ~ coef_var + (1/study)
Highest NPV	diagnosis ~ IQR
Highest Average Across metrics	diagnosis ~ coef_var + (1/study)

Measure	Model
Most accurate model	diagnosis ~ coef_var + (1/study)

Comparing random effects used

Models with study as RE as well as models without RE at all perform much better than the rest of the models. Gender generally performed poorly, except in combination with study.

Figure 13: Performance of single predictor models



After excluding the worst performing models (Based on the plot): * study + ID * ID * ID + Gender

Which single predictor performs best? Following plot shows that our models are generally more sensitive than they are specific. A high sensitivity means that we have a low number of false negatives (i.e., people with schizophrenia not diagnosed).

However, the lower specificity means that we do have a considerable amount of false positives (i.e., people without schizophrenia who are diagnosed as having schizophrenia).

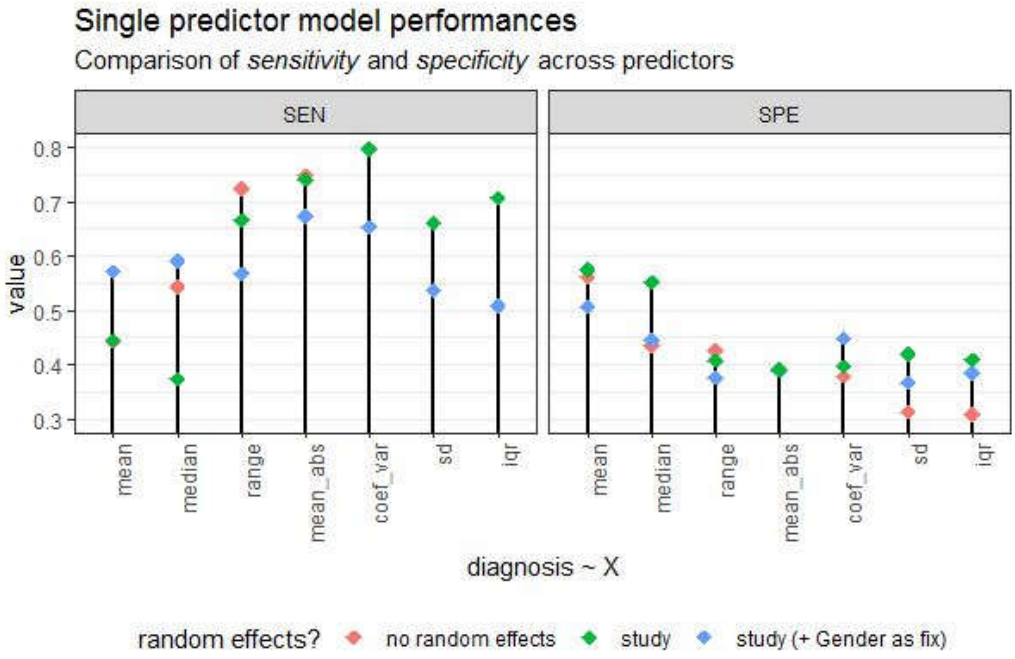
Trade-off between sensitivity and specificity: It is important to be alert to what the real world consequences of making different classification errors is. We would rather incorrectly label a non-schizophrenic person 'schizophrenic' than mislabelling a schizophrenic (taking the viewpoint of the nurse, not the local government).

With that said, the model 'diagnosis ~ iqr' seems quite extreme as it has more false positives than true negatives (SPE). As can be seen in the plot, although it scores very high on the sensitivity score, it is the lowest scoring single predictor when it comes to the specificity score.

The additional issue with low specificity is that the base rate of schizophrenia in the general population is lower than in our sample, so this model would produce way too many false positives (as is the issue with breast cancer screening). The best model (as suggested by the table above) is the one with coef_var as predictor.

The NPV & PPV values support the fact that either coef_var or interquartile range (IQR) is the best predictive model that we have.

Figure 14: Single Predictor Model Performances



IQR (with no random effects - the red dot) does slightly better on NPV, while coef_var (with study as random effect) does slightly better on PPV.

The NPV value reflects how likely it is the someone diagnosed as a negative (not having schizophrenia) actually does not have schizophrenia.

The PPV value reflects how likely it is that someone diagnosed as a positive (having schizophrenia) actually does have schizophrenia.

This brings us back to the dilemma between the nurse (caring mainly about the patient) & the local government (caring mainly about money).

To see what is happening, we also made a plot showing Brier scores across predictors.

The best model (on average) is the model with coef_var and study as random effect (green dot).

The coefficient of variation expresses the dispersion of a distribution and is computed as the ratio of the standard deviation to the mean. In this case, it might be the best predictor because it takes both the mean and standard deviation into account.

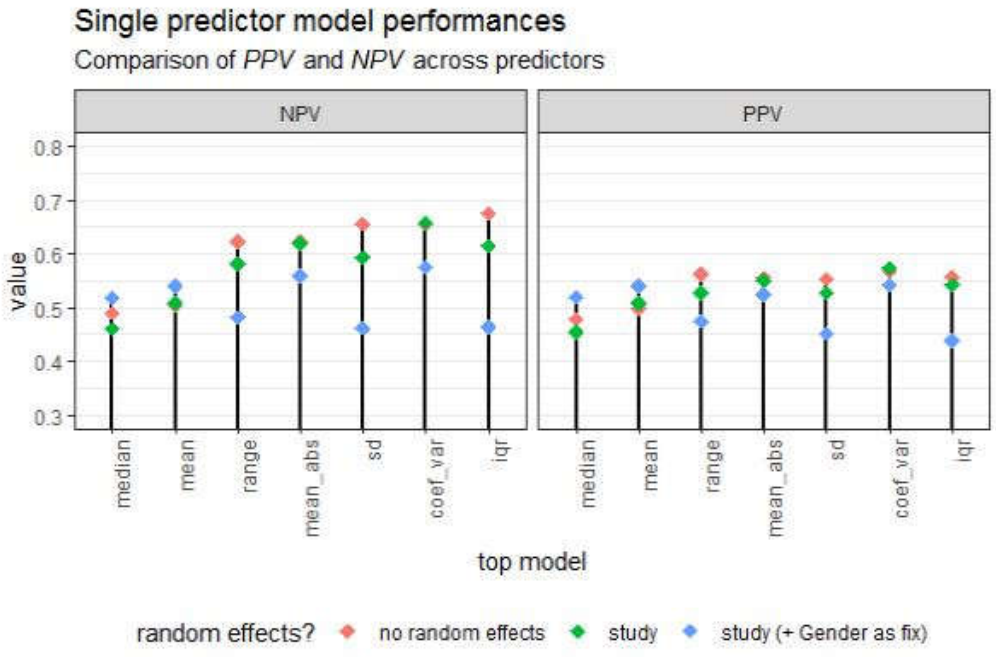
Task 3, Best combination of predictors

In order to find the best combination of predictor variables, we combined all the predictor variables and random effects in all possible ways. This gives us a total of 1.533 possible models which we cross-validated.

Metric	Model
Highest Sensitivity	Diagnosis ~ IQR
Highest Specificity	Diagnosis ~ mean + range + IQR + (1/study)
Highest PPV	Diagnosis ~ mean + sd +range + mean_abs + se + Gender + (1/study)

Metric	Model
Highest NPV	Diagnosis ~ mean_abs + coef_var + se + Gender
Highest average metric	Diagnosis ~ mean_abs + coef_var + se + Gender

Figure 15: Single Predictor Model Performances



Picking a model

Because the primary objective of this part of the assignment was to construct the best predictive model, we decided to forego the explanatory approach. Instead, we decided to base our selection of the "best" model solely on model performance as measured by brier scores.

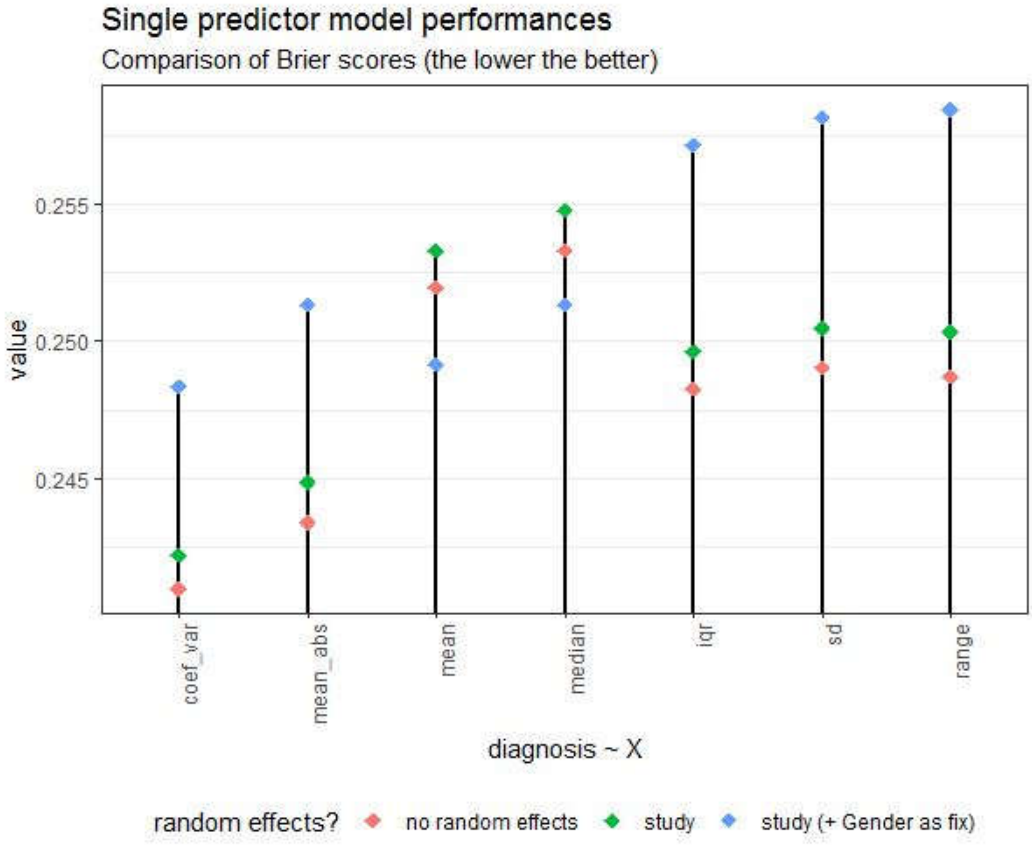
A brier score measures the accuracy of probabilistic predictions, applicable when the outcome is binary and mutually exclusive.

The brier score is calculated by subtracting the actual outcome from the predicted probability of the outcome, which is of course a binary value. A lower brier score indicates accurate model performance.

Taking this into consideration, we ended up with four (quite similar) models with highly similar brier scores.

Model	Brier Value
Diagnosis ~ coef_var + se + Gender	0.2293
Diagnosis ~ coef_var + se + (1/study)	0.2295
Diagnosis ~ coef_var + se	0.2296
Diagnosis ~ IQR + coef_var + se + Gender	0.2298

Figure 16: Single Predictor Model Performances



Evaluated strictly in terms of lowest (best) brier value, we choose model 1 as our best model. The model formula looks like this.

`diagnosis ~ coef_var + se + Gender` , with a brier value of 0.2293.

So does the model actually have any real-world applications?

Diagnosing schizophrenia can be done based on voice, but the accuracy of doing so leaves something to be wanted. Other information related to the participant, should, if possible, be taken into consideration.

Given a clinical setting one might choose to weigh the different accuracy measures depending on priorities. If one cares less about mislabelling non-schizophrenic people and more about "catching" schizophrenics, one would weigh sensitivity and NPV higher than specificity and PPV.

However, if we are obliged to predict schizophrenia based solely on voice, our model offers better performance than the standard approach of simply taking the mean or the standard deviation.