Universiteit Utrecht

Differentiating psychotic patients by linguistic features

Clustering patients with psychotic disorder to explore the relationship between diagnostic and linguistic properties

P.W. Barkema - 5979412
pieterwbarkema@gmail.com

Artificial Intelligence
Utrecht University

Supervisor: Dr. H.G. Schnack

Bachelor's thesis 7.5 ECTS

Abstract

Psychotic disorder causes high social costs due to the impact it has on patients and the high prevalence, especially among adolescents. No reliable biological indicator exists for the diagnosis of psychotic disorder, although research shows language has potential to become a biomarker. One symptom of psychotic disorder is incoherent language. In this research paper the use of incoherent language to differentiate between different groups of patients was explored. Incoherent language was represented by a feature set extracted from the interviews of 50 patients and 50 healthy controls (N=100) processed with word2vec semantic analysis. Features were chosen by their ability to separate patients from controls. We then used those language coherence features to group psychotic patients using unsupervised clustering. Multiple cluster models successfully clustered the patients with up to four features. The general symptom score was significantly different between clusters and no confounding factors were found. This exploration shows the usefulness of clustering techniques for this particular use case. It is among the first evidence that symptom severity measures of psychotic disorder and linguistic coherence may be related. This could be the first step towards the detection of illness severity by language coherence, which could help provide timely care for the patient.

August 13, 2019

Contents

1 Introduction

Psychotic disorder is a very debilitating mental illness, characterized by a group of symptoms called psychosis. Sufferers rarely manage to maintain, let alone build up, a normal life. Psychosis covers many varieties of episodical delusions and sensory hallucinations and may be the harbinger of a psychotic illness as schizophrenia, for example. Schizophrenia is known for its high societal costs due to its high prevalence, especially among young adults [Delespaul et al., 2013]: almost one percent of the world population is estimated to be schizophrenic. Schizophrenics have a 15 year lower life expectancy [Crump, 2013] and research found those affected by schizophrenia are only employed in 20% of instances, and in the US one-fifth lost employment within a year of a first psychotic episode. Schizophrenia is one of nine possible diagnoses within the psychotic disorder spectrum. These statistics have been attributed to lack of fundamental understanding of the disease and, implicitly, lack of appropriate tools for diagnosis [Insel, 2015]. Psychotic disorder is thought to be more than just a chemical imbalance in the brain [Insel, 2015] [Sommer, 2017].

This lack of understanding calls for a more modern approach. Originally, five versions of the Diagnostic Statistical Manual of Mental Disorders (DSM) have provided us with five different definitions of psychotic disorder inspired by the focus of psychiatric research at the time [Marquand et al., 2016]. The DSM is the international standard for psychiatrists. It aims to provide symptom-based checklists for diagnosis and consensus about terminology. However, DSM diagnosis costs much time, lacks reliability and causes heterogeneity within diagnostic categories [Wang and Krystal, 2014]. A newer approach for description and detection is computational psychiatry. This field uses data-driven solutions to unravel categories of mental disorder without these disadvantages. What data should be used to detect psychotic disorder is unclear, however. Reliable biological markers are unavailable, but have potential [Insel, 2010]. Certain different methods, such as fMRI and genetics analysis, have been useful, but none have accurately predicted or diagnosed the illness. One reason could be that biological validation is not the intention of the DSM's treatment-oriented definition of psychotic disorder [Insel, 2014]. Another problem is the sample size needed to verify biomarkers. The groups of patients and healthy controls need to be both homogeneous to prevent overfitting and large enough in size so convincing results for the disease can still be found [Schnack, 2017].

Natural language has potential as a biomarker and could help provide better diagnostic tools. Natural language entails the structure and meaning of spoken language. This should not be confused with speech, which focuses on acoustic information rather than content. A common symptom of psychotic disorders is abnormal language. Sufferers from the illness may have their language impacted syntactically and semantically. Syntactically, a plethora of research backs up that psychotic patients use simpler syntax, such as shorter sentences [Corcoran et al., 2018]. Semantically, psychotic disorder is often accompanied by disarray in language. In serious cases patients might leap from subject to subject quickly or even tentatively: without ever returning to the original subject [Iter et al., 2018]. Kraeplin already noted in 1913 that patients suffer from Sprachverwirrtheit (Kraeplin, 1913, as cited in Corcoran et al, 2018). Ever since, a large body of research has been dedicated to the relation between language coherence and psychosis. Despite recent progress in computational solutions experts say the patterns of this distorted language are still poorly understood [Corcoran et al., 2018]. This symptom can often lead to difficult conversations with patients and forms an obstacle for diagnosis. For the particular case of psychosis this could be turned into an advantage. Because language incoherence is so characteristic for psychosis, it could be quantified to become a reliable biomarker for psychosis [Sommer, 2017]. Some attempts to accomplish this have been made, but without undisputed success.

State-of-the-art research attempts to bridge the research gap between psychosis and language with models for detection and quantification of linguistic incoherence. Different concepts have been applied, such as speech graphs [Mota et al., 2012] and Latent Semantic Analysis [Bedi et al., 2015]. Research analyzing the latent semantics of patients with word2vec has been led by Rezaii, Walker & Wolff (2019). Word2vec models convert words to vectors that represent the chances words co-occur in context. Training the model on large corpora of coherent spoken language could train it to detect abnormal co-occurrences in language of psychotic patients. The model creates a semantic space of words in training data, which can be used to calculate how quickly a person covers distance within this space when speaking. A subject speaks more incoherently when they use sequences of words that are uncommonly used together in training data, which leads to large leaps within a short time in semantic space. This approach in particular has been praised for both empirical success [Voleti et al., 2019] [Altszyler et al., 2016] and potential [de Boer et al., 2018].

The aim of the research is to explore the relationship between diagnostic scores for measuring psychotic disorder and linguistic coherence. How linguistic coherence should be defined is unclear. The language coherence measures used in literature should be used with caution, because they have not been proven to accurately work yet. We propose our own coherence measures engineered from quantified language of patients and validate them. These measures will be inspired by literature and aim to represent linguistic coherence. Current endeavours mostly focus on differentiating controls from patients by language, even though mental illness is widely known to be a spectrum rather than a binary class [Adam, 2013]. Little research in this field has been done to distinguish categories within patients. Language coherence comparison between patients is essential for understanding the relationship between language coherence and psychotic disorder. We will therefore attempt to explore this relationship by identifying groups with different linguistic coherence properties. The final goal is to investigate whether patients with different linguistic coherence scores have different diagnostic properties, as well.

The three research questions are summarized as:

- What set of features can be used to represent language coherence?
- Can coherence feature be used to distinguish different categories within patients?
- What different diagnostic properties do the resulted categories of patients have?

To the first question we suggest a small set of coherence features will be validated. When researching the second question we expect to find at least two categories with different linguistic characteristics in the data. Due to the relationship between linguistic coherence and psychotic disorder and the fact that psychotic disorder is a spectrum, we expect to see diagnostic differences between the groups of subjects. Linguistic incoherence, like other disorganizing symptoms, is known as a positive symptom of psychosis. Consequently, in the third phase of the research we expect positive symptom scores will differ between distinguished groups.

The methodology of research will be shortly described. From 50 patients and 50 controls (N=100) interviews about general topics were held previously to this research. Word2vec was then used to convert these interviews to vectors as suggested by most literature. Afterwards, the cosine similarities of a sliding window over these vectors were calculated to provide raw coherence data. We now propose a set of nine features inspired by literature, consisting of standard statistical features and time-series features. These coherence features will be validated by their ability to separate controls from patients. The validated features are assumed to represent coherence. Secondly, a method is needed that groups data points with relatively similar coherence features together. The algorithm should not be guided by specific DSM diagnoses or scores, introducing bias [Brodersen et al., 2014], but rather define its own groups. Because it is explorative research, we do not want to assume these labels to be the ground truth. Only the label psychotic disorder is used as it is being researched. Unsupervised learning is suited for this objective. The main edge of learning without a teacher is that the algorithm can identify its own classes based on an underlying structure. Unsupervised clustering techniques can find the best partition of the patient data based on the validated coherence features. Techniques as k-means clustering and Hierarchical Clustering can be used for this purpose. The best models can be selected with model evaluation techniques. The linguistic differences between the resulted clusters can then be interpreted. Statistical techniques can be used to test for diagnostic score differences between members of different clusters and to find factors confounding the relationship between linguistic and diagnostic differences.

2 Methodology

2.1 Participants

The data consists of 100 Dutch participants who participated knowingly in the collection of data for research purposes. The 50 subjects have been diagnosed with psychotic disorder according to DSM V and 50 subjects are healthy controls. The subjects are aged between 18 and 73 years. 74% is male and 26% is female. Controls with known (family) history of mental health issues were excluded, as well as subjects with speaking or hearing disorders. Only Dutch speakers were included. Informed consent was taken from all participants. They received compensation of 10,- euros.

2.2 Materials

An AKG-C544l head-worn cardioid microphone was used to record the subject's speech. Speech was digitally recorded onto a Tascam DR40 solid state recording device at a sampling rating of 44,100 kHz with 16-bit quantization. The language coherence data was quantified with the gensim word2vec module for word embedding. The model was finetuned at $dimensions = 300$, $minimum_occurences = 5$ and $window_size = 9$. The process of finetuning has been done in previous research [Voppel et al., 2019] and is outside the scope of this paper.

All programming code was written in Python (version 3.7.0) in the Spyder IDE from Anaconda. The data processing and modelling techniques used were imported from the Machine Learning library sklearn and the statistics modules for tests and feature extraction from scipy.

2.3 Data collection

All data had been collected before this research. They were processed and supplied by the University Medical Centre Utrecht. The data consist of raw language coherence data, diagnostic scores, demographic features and medical information.

2.3.1 Language coherence data

The language coherence data consist of one numerical vector per subject with values between 0 to 1. This was produced from text data in three steps (Figure 1). First, an interview was conducted for 5 minutes per participant in Dutch. The interview aimed to gather as much participant speech in this time frame as possible. The interviewer asked several general personal questions so any subject had the chance to answer elaborately. Once a subject was done speaking, another question was asked. An example of such a question is: "what was your first swimming lesson like?"[1].

Seven contributors, researchers and interns, transcribed the interviews in double blind conditions. All contributors transcribed at least one patient and one healthy control. Per intern at least one transcription was double-checked for quality by a researcher.

Secondly, word2vec [Mikolov et al., 2013] was used. This is a neural network which converts words to vectors using the context they are used in. The model trains by iterating over large corpora of text with a sliding window to create a vector representation of every word. The vector contains the chances of co-occurrence within a specified distance with every other word. The model learns to produce vector representations compressed to a specified number of dimensions d. These vectors altogether form a semantic space as visualized in Figure 2. Interesting relationships can be found in this vector space. Word2vec for instance fills in the gap 'Greece' relates to 'Athens' as X to 'Istanbul' with 'Turkey'.

Because the model is sensitive to specific words, the train set should be similar in language, time period recorded and data type [Yang et al., 2018]. Languages are complex, so large amounts of data are needed to properly account for all the necessary contexts of all words. The word2vec model was trained on the largest corpus of spoken Dutch language available: Corpus Gesproken Nederlands (CGN) [Oostdijk, 2000]. Only the documents containing spontaneous language were

[1] A full list of the questions is available online at Schizophrenia Research

used to create a more reasonable representation of unprepared language.

Thirdly, the vectors that the trained model had produced per interview were grouped. They were divided into subsets of vectors to catch differences in coherence throughout a text to allow for more complex comparisons. The text could not be divided into sentences, however. Sentence boundaries are hard to define in spoken language. Instead, each interview was subdivided in windows. Another sliding window was introduced for this, which came in two flavours. The summary modus produced a partitioning of the text into windows of equal size. The simple modus extracted windows of size w. With simple mode the first window contains words with index 1 to w. Then, the index is increased by one and the next window contains words with index 2 to $w + 1$. This continues until the window reaches the final word. The window coherence data for both modi was extracted for window sizes 2 through 20. The cosine similarity of all windows was calculated to produce the language coherence data. The window coherence data will be used to compare the coherence of different interviews.

Lastly, the coherence windows were compressed to singular values. A window is coherent when the words are often used in the same context together, causing the words to become similar vectors. The used similarity measure is cosine similarity. The coherence of a group of words is then defined as the cosine similarity of the corresponding vectors. The cosine similarity calculates the cosine of the angle between vectors. For a vector v_1 and vector v_2 with v_i as the ith element of a vector v the cosine similarity is calculated as:

$$\cos(\mathbf{v_1}, \mathbf{v_2}) = \frac{\mathbf{v_1} \cdot \mathbf{v_2}}{\|\mathbf{v_1}\|\|\mathbf{v_2}\|} = \frac{\sum_{i=1}^{n} \mathbf{v_1}_i \mathbf{v_2}_i}{\sqrt{\sum_{i=1}^{n} (\mathbf{v_1}_i)^2} \sqrt{\sum_{i=1}^{n} (\mathbf{v_2}_i)^2}} \tag{1}$$

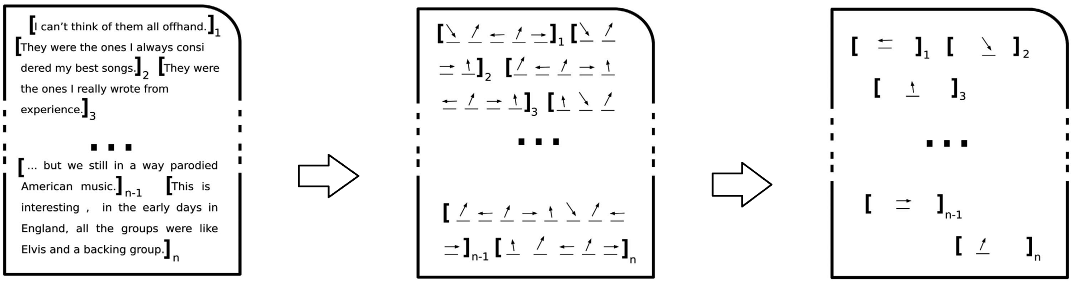

[I can't think of them all offhand.]1
[They were the ones I always considered my best songs.]2
[They were the ones I really wrote from experience.]3
...
[... but we still in a way parodied American music.]n-1
[This is interesting , in the early days in England, all the groups were like Elvis and a backing group.]n
1
2
3
...
n-1
n
1
2
3
...
n-1
n

Figure 1: The raw coherence data production steps applied to sentences instead of windows ([Bedi et al., 2015])

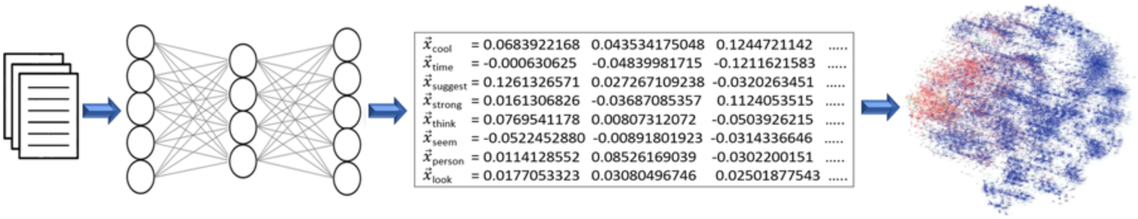

$\vec{x}_{cool}$ = 0.0683922168 0.043534175048 0.1244721142 .....
$\vec{x}_{time}$ = -0.000630625 -0.04839981715 -0.1211621583 .....
$\vec{x}_{suggest}$ = 0.1261326571 0.027267109238 -0.0320263451 .....
$\vec{x}_{strong}$ = 0.0161306826 -0.03687085357 0.1124053515 .....
$\vec{x}_{think}$ = 0.0769541178 0.00807312072 -0.0503926215 .....
$\vec{x}_{seem}$ = -0.0522452880 -0.00891801923 -0.0314336646 .....
$\vec{x}_{person}$ = 0.0114128552 0.08526169039 -0.0302200151 .....
$\vec{x}_{look}$ = 0.0177053323 0.03080496746 0.02501877543 .....

Figure 2: The word2vec pipeline goes from text through trained neural network to semantic space [Rezaii et al., 2019].

2.3.2 Diagnostic information

The diagnostic data consist of positive, negative, general and total symptom scores as defined by the Positive and Negative Symptom Scale (PANSS) scores [Kay et al., 1987]. Positive symptoms add processes, such as delusions, while negative take them away, such as apathy. Other symptoms, like anxiety, fall in the general category. The total is the sum of these three scores. The PANSS score is measured by trained interviewers who fill in three separate checklists for 7 positive, 7 negative and 12 general symptoms, scoring every symptom for a subject from 1 to 7. The respective ranges of the PANSS scores are negative [7..49], positive [7..49], general [16..112], total [30..210]. The PANSS interview was administered by two qualified researchers in the study the patient had partaken in. The PANSS score was calculated as the consensus score of the two interviews. The 50 patients on average have a total PANSS score of 53, PANSS positive score of 11, PANSS negative score of 14 and PANSS general score of 27.

2.3.3 Demographic information

The demographic features were included in the research, because they could be confounding incoherent language. For example, older adults are "more likely to produce tangential, off-topic utterances in conversation" and could produce less coherent answers [Hoffman et al., 2018]. It is also thinkable that more education leads to more coherent speech. The demographic information of all subjects was gathered using the CASH questionnaire. The demographic features used were education (years) (YOE), average education of parents (years) (YOEP), age (years) and gender (male/female). The mean and distribution of the features are displayed in Table 1. The table also contains the results of the non-parametric Mann-Whitney U test for significant difference between patients and groups per feature (significant values in bold with $\alpha = 0.05$). The YOE is defined as the years of education from primary school onwards a subject had experienced. The YOEP is defined as the years of education from primary school onwards both parents of a subject had experienced, averaged between them.

Table 1: Demographic features of subjects (N=100).

	Patients (N = 50)	Controls (N = 50)	Patients compared to controls (p-value)	Subjects (N = 100)
Gender (% male)	64%	84%	.1853	74%
Age (years)	29.2000 (σ = 9.0178)	31.4200 (σ = 12.3062)	.3072	30.7323 (σ = 10.8450)
Education (years)	12.8571 (σ = 2.7553)	14.4800 (σ = 2.3853)	**.0014**	13.6768 (σ = 2.6999)
Average Education Parents (years)	12.2955 (σ = 3.1176)	12.6122 (σ = 2.9970)	.2985	12.4624 (σ = 3.0588)

2.3.4 Medicine type information

Medicine information was used to compare between different groups of patients. 45 out of 50 patients used 9 antipsychotic medicine in total. Antipsychotics were divided into two categories according to the amount the active ingredients bind to dopamine D_2 receptors (D2R). Medicine that bind to D2R in any grade are known to inflict disturbances in speech. These disturbances could increase as the grade increases. The medicine will be divided into high D2R occupancy and low D2R occupancy [de Boer et al., 2019]. Patients that do not use antipsychotics are added to the low category, because of the low amount of data points. Medicine data were included to check if medicine use could be confounding results. There is no evidence for this. However, research has found that methylenedioxymethamphetamine (MDMA) and methamphetamine have affect the semantic characteristics of spoken language [Bedi et al., 2014]. These substances also bind to the dopamine receptors. The mentioned semantic characteristics are very similar to the ones used in this research.

High D2R medicine = $[''amisulpride'', ''aripiprazol'', ''fluanxol'', ''haloperidol'', ''risperdal'']$ Low D2r medicine = $[''clozapine'', ''olanzapine'', ''paliperidon'', ''paliperidondepot'', ''quetiapine'']$

2.4 Coherence feature extraction

Every subject's interview has been split up into windows and processed to a set of cosine similarities from which features can be extracted. Several different types of features will be proposed to try capture the linguistic properties per subject. There was little validated literature in this field to base our choices on. Bedi et al. (2013) divided text up in sentences and used the *minimum* of the cosine similarities per text as a feature. This was effective, but their low sample size combined with the complexity of their model makes the generalizability of the model questionable. We propose a set consisting of features based on literature and general purpose measures, alongside several explorative features. This was also based on experience with the data set, statistical tests and expected linguistic properties. We propose nine features to be extracted per interview.

2.4.1 Standard features

Some basic statistical measures are used as features. The arithmetic *mean*, *median*, *variance*, $standard_deviation$ (for its standardized properties), *minimum* and *maximum* were extracted for every one of 100 sets of cosine similarities.

2.4.2 Explorative features

We propose the feature $difference$ to calculate how much the cosine similarity coherence measures swings throughout the interview. This quantifies the differences between adjacent coherence windows. **Intuition:** The idea is that a more incoherent speaker could maintain a steadier or more turbulent coherence level than the patient. For every window and its neighbouring window the absolute difference in cosine similarity is summed up. The $difference$ measure for the set of coherence windows' cosine similarities cw of size d is defined as:

$$difference = \frac{\sum_{i=1}^{d-1} ||cw_i - cw_{i+1}||}{d-1}$$

Now we introduce two time series features. An interview's coherence could be inherently different very early in an answer as opposed to later, for example. The linguistic coherence may depend on the structure of the overall story. With this idea of coherence being time dependent we could attempt to use time series features. Two time series features from literature [Laptev et al., 2015] will be proposed to represent coherency development within a subject's interview.

Measure of skewness

Intuition: The idea is that patient coherence data might contain relatively more low values. The data might be skewed to the incoherent (left) side of the mean. *skewness* measures the (lack of) symmetry of the data distribution compared to the normal distribution. This feature measures how long the tail of the distribution is. A tail that is longer or fatter on the right has positive *skewness*. This also means that the median and mean are greater than the mode. Patients are expected to have higher *skewness* due to higher frequencies of low coherence windows. *skewness* for N coherence windows with value Y and sample standard deviation s is defined as:

$$\mathbf{skewness} = \frac{\sum_{i=1}^{N} (Y_i - \bar{Y})^3 / N}{s^3}$$

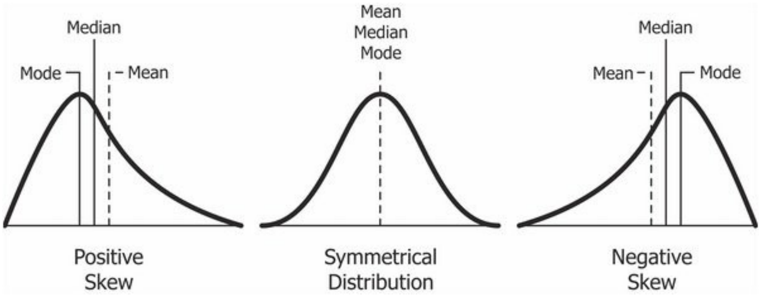


Figure 3: Positive, neutral and negative *skewness* visualized with the mean, median and mode labels.[2]

Measure of kurtosis

Intuition: The idea is that unhealthy subjects might have more extremer swings in coherence, and defer more extremely from their mean coherence compared to healthy subjects. *kurtosis* measures the peakness of the data compared to a normal distribution by comparing both tails. Distributions with longer tails on both sides have many outliers and low *kurtosis*, while distributions with small tails and few outliers have high *kurtosis*. Patients are expected to have more outlying coherence values and thus a lower *kurtosis*. This does not reflect the same property as standard deviation, but rather focuses on the degree of extreme values in a distribution. Two distributions can have the same standard deviation with different distribution shapes and different tails. *kurtosis* for N coherence windows with value Y and sample standard deviation s is defined as is defined as:

$$\mathbf{kurtosis} = \frac{\sum_{i=1}^{N}(\mathbf{Y_i}-\bar{\mathbf{Y}})^4/\mathbf{N}}{\mathbf{s^4}}$$

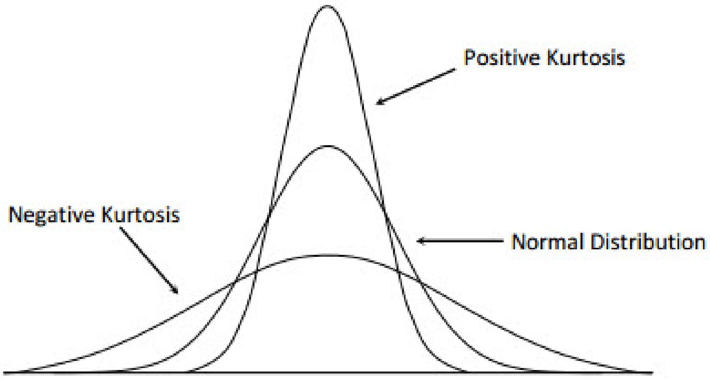


Figure 4: Positive, neutral and negative *kurtosis*.[3]

2.4.3 Dimensionality reduction

The raw coherence data per interview cannot directly be used as features, because the ratio of features to data points is disproportionally high and it likely contains large amounts of noise. There are hundreds or thousands of windows per subject and only 100 data points. This 'curse of dimensionality' causes the model to generalize poorly. Regardless, it is desirable to retain as much information as possible, but our hand-picked features throw out much information. Principal Component Analysis (PCA) is a dimensionality reduction technique. It can compress a large data set to d dimensions while retaining as much information of the original data as possible. It does so by calculating the eigenvectors of the features' covariance matrix. This entails that in a point cloud PCA extracts the d main vectors (figure 5[4]). These vectors then explain as large of a percentage of the data's variance as possible. d is ideally tuned until enough information is retained while still dramatically reducing the amount of dimensions. The extracted vectors are named the principal components and can be used as features, because they contain similar information to the original data. Even though some information may be lost, this would be compensated for by the reduction of dimensionality and noise reduction.

A rule of thumb states that the amount of components should at least explain > 90% of the variance in data. PCA requires a consistent feature set size for every data point. However, the data points are all of variable size, because of the different interview lengths. With a cut-off at the 100th raw coherence measure (the lowest data set size) PCA was tested and required 47 components to explain > 90% variance. Because the interviews varied greatly in length, most data is thrown away at the cut-off. Moreover, it only reduces the dimensions by half, while only retaining just over 90%. This poor performance was partially expected. PCA mainly works well for (highly) correlated values, because correlation brings redundant information. It is very likely, and apparent

[4] Source: *https : //en.wikipedia.org/wiki/Principal_component_analysis*

from the PCA results, our raw coherence measures are not correlated. Due to the bad test results, the principal components will not be further analyzed.

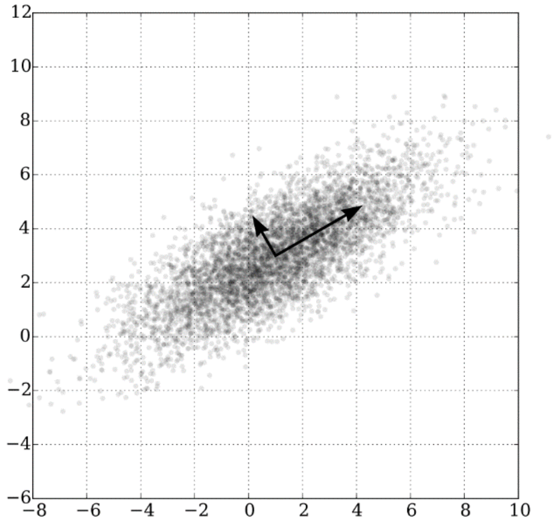


Figure 5: The eigenvectors of a point cloud are used to extract principal components for PCA.

2.5 Preprocessing

2.5.1 Feature selection and validation

These proposed features aim to represent language coherence. After feature extraction the data set contains 9 features for every window size 2 to 20 per two window modi per data point. The aim is to reduce this feature set to ensure there is enough data to generalize our findings. This narrows our search and removes redundant information. Firstly, we will decide what window type makes the most sense for this research, then the best window size per feature will be selected. The simple window modus was chosen, because it outperforms summary mode in all basic performance tests and makes more sense. The partitioning summary mode arbitrarily picks the contents of coherence windows and creates much less windows. The simple mode's moving window creates a coherence measure for every combination of possible neighbours for a word. Perhaps, summary mode would have a speed and memory advantage in large corpora, but with one interview per subject, the simple mode is more informative. Moreover, recall that the word2vec model, which created the raw coherence data, was trained using a sliding window. Perhaps, our model will consider word windows more similar to word2vec's word windows with the same modus. The simple modus will be used to process all interviews throughout the research.

The smallest windows (2 to 4) and largest windows (16 to 20) will not be taken into consideration to limit the effect of outliers and topic transition noise. Topic transition noise is caused by windows, which overlap over multiple interview answers. This causes the coherence in between answers to be calculated, which is not the purpose of this research. This is a side effect of using sliding windows. The upper and lower threshold of the window sizes do not need to be precise. When the size of the simple mode increases by one the windows only contain one word extra. They contain nearly the same information, allowing a margin of error.

One window size per feature is chosen to test the potential of the feature in distinguishing patients and controls. The assumption is that whatever window size maximizes the difference between patients and controls is the best fit for a feature to represent coherence. The validation is done by testing if a coherence feature has the same distribution with mean μ_1 for patients as the distribution with mean μ_2 for controls is tested. The H0 hypothesis: $\mu_1 = \mu_2$ and H1 hypothesis H1: $\mu_1 \neq \mu_2$.

A coherence feature is considered validated if the feature as a property of all patient interviews is significantly different from the control interviews (significance level $\alpha < .05$). Those measures of coherence are then shown to be different between groups. Not significantly different features can still be informative for the model. From the feature set the most informative ones will be distilled with model selection. These features can then be used to distinguish different classes of linguistic coherence within patients.

2.5.2 Missing data and outliers

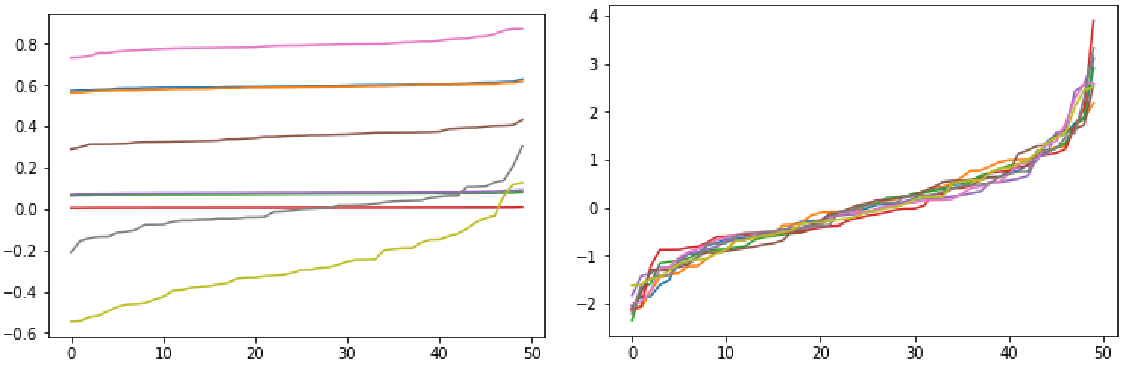
All subjects with missing values will be assessed. From the data set 8 subjects have missing demographic information, and controls have no diagnostic score by definition. The 8 subjects will be included in feature selection, because their coherence and diagnostic information is still intact. They will only be excluded from the demographic analyses of the clusters where their information is missing. The 50 controls will be used for feature selection, but by definition have no diagnostic scores. They cannot be compared with patients and thus excluded from cluster analysis. The data was previously vetted well and invalid or outlying data had already been removed. Therefore, no outlier detection will be done beforehand, since there will be no basis to remove data on if outlying values are detected. Cluster models may indicate noisy data points, which will be removed.

2.5.3 Normalization and standardization

Cluster techniques can be sensitive to scale differences, depending on the distance measure used. The general use case Euclidean distance will give privilege to higher absolute differences. A clustering technique will be making the invalid assumption that *kurtosis* and *mean* are of comparable scale when describing the difference between two data points. Therefore, all features will be standardized to account for the difference in feature standard deviation and mean between features. Every feature will be standardized to have a mean of 0 and unit variance ($\sigma = 1$). Standardization of an array of feature F with mean μ and standard deviation σ to F_s means that every standardization of $f \in F$ with length N to $f_s \in F_s$ is defined as:

$$f_s = \frac{f-\mu}{\sigma}$$

In figure 6a and 6b the effect of standardization is visualized. It appears to be normally distributed.



(a) All features visualized unstandardized.

(b) All features visualized standardized.

Figure 6: The effect of standardization on the feature set.

3 Model description and evaluation

3.1 Clustering methods

Unsupervised clustering methods group data points based on their relative similarities. There are many flavors of cluster methods, using different algorithms and distance measures. All can be used to find underlying patterns in data to divide data sets into clusters. These underlying patterns can be found within feature (dis)similarities. Interpretation and characterization of these clusters are often complex processes, because the process is unsupervised. There is often not a crystal clear conclusion to draw from the resulting clusters, but clustering rather hints at certain multivariate effects that distinguish groups of data points.

The data set is too small to use complex clustering models. There is also no relevant literature to suggest a problem-specific clustering model. The techniques will be limited to two general use case cluster models. Model selection will be performed with k-means clustering and Hierarchical Clustering. Density-based models, such as DBSCAN, are not expected to work, because they assume similar density of clusters. Basic tests have not shown this is the case. A short summary of both algorithms is provided below.

3.1.1 K-means clustering

K-means clustering divides the data set into a predefined k amount of clusters. It does not detect noise. K-means follows the following steps. Given a set of data points, choose k random data points to be the cluster centers or centroids for the first stage. Assign every data point to the closest cluster center, and recompute the centroids to be the center of the data points currently in that cluster. The algorithm has not converged if data points are still switching memberships. Then, recompute the centroid and reassign memberships to the data points accordingly until it has met the convergence criterion. We will implement random restart for k-means to account for the random choice of initial centroids, which causes local optima to be found. The amount of random restarts will be increased until the validation score converges. The model that uses the k with the highest validation score becomes final model.

Strengths

- It is easy to understand and implement.
- Time complexity is only $O(n*k*i*d)$ [Wu et al., 2008] with n: number of points, k: number of clusters, i: number of iterations, d: number of features. k, i and d are generally small integers, so it runs roughly in linear time.

Weaknesses

- It terminates at a local optimum if sum squared of errors (SSE) is used for model selection. The global optimum is hard to find due to the random choice of centroids.
- The user needs to specify k.
- The algorithm is sensitive to outliers due to its inability to detect noise. It will force them into a clusters, which will affect the centroid's location.

3.1.2 Hierarchical Clustering

Hierarchical Clustering (HC) creates a hierarchy in data based on distance. A linkage algorithm uses some distance measure to calculate the distance between all data points. HC then has two algorithms to process this linkage. Agglomerative clustering will be used throughout this paper, but divisive clustering will be mentioned for completeness.

Agglomerative clustering: the dendrogram is built from the bottom level. The algorithm starts with a set of singleton clusters as the root. These merge recursively with some nearest cluster calculated by applying a specified merging criterium to the linkage distances. The nearest cluster could be defined as the nearest on average or the cluster with the nearest cluster member (minimum distance), for example. In this research we only use the group average linking criterium, because we found no reason other criteria would work better. This criterium finds the closest cluster by averaging all data points of clusters, weighted by the amount of members in the cluster. This process continues recursively for every cluster until one cluster remains.

Divisive clustering: the cluster is built from the top level. The algorithm starts with all data points in one cluster as the root and then recursively splits it into two child clusters by applying some merging criterium to the linkage distances. It then recursively divides the children clusters until only singleton clusters remain.

Both algorithms generate a dendrogram, which visualizes all merges or divisions in order with the relative distance bridged per recursion. When the marginal gain becomes too little, the process should be stopped. The best clustering can be found this way.

Strengths

- It provides the insightful dendrogram, making retracing easy and indicating possible outliers.
- It provides an intuitive algorithm with possibly different results from k-means.

Weaknesses

- It involves lots of arbitrary decisions, such as the merging criterium, distance measure, and ultimately the chosen cluster size, although there are heuristics for this.
- Pairing or division of data points cannot be undone. The algorithm will always group the same data points together: Once the damage is done, it can never be repaired. (Kaufman, 1990, as cited in Hintze, 1992).
- The algorithm does not scale well. The naive algorithm runs on time complexity $O(n^3)$ and memory complexity $O(n^2)$ due to the linkage algorithm. There are some use cases with better complexities.

3.2 Model evaluation

The cluster techniques we use will always yield some result. Unfortunately, this does not mean the result is informative at all. Even a data set that completely consists of noise will be clustered by k-means and HC. Model evaluation will be done through heuristics, such as cluster validation measures and highest memberships similarity, to guide model selection. Although these are good indications, a priori knowledge is more important. If a best model divides a heterogeneous group of 50 subjects into 20 clusters, that is not very plausible. That is why solution quality is used as a subjective element to judge the outcome.

3.2.1 K-means validation

One objective error measure for cluster validation is the sum of the squared distances (SSD) between data points and their corresponding cluster centroids. This measure is unreliable, however. Adding clusters will always lower the total SSD. Naturally, with N data points and $k = N$ the SSD is 0, because every data point is its own centroid. Removing features also lowers the total sum of distances between data points. However, the marginal profit in distance per added cluster can be used to pick the best k. A rule of thumb known as the 'elbow rule' suggests that the marginal decline of distance per added cluster drops so much at one point that adding another cluster is not worth it. This drop looks like an elbow on the SSD per k curve. Of course, this threshold is both arbitrary and ambiguous (figure 7). There is not always a clear elbow and the measure does not take in account the quality of the clusters.

An objective validation measure which translates across any number of features and clusters would be ideal. The average silhouette coefficient scores a model between -1 to +1 comparing the average distance of a data point to its assigned cluster compared to the nearest other cluster. A score of zero indicates a data point is as similar to its current cluster as to the closest neighbor. Silhouette score s [5] for a sample is defined as:

$silhouette(s) = \frac{b-a}{max(a,b)}$

a = the mean of distances from a sample to all points in its assigned cluster

b = the mean of distances from a sample to all points in the nearest other cluster

The average of the silhouette coefficients for all data points is a good validation measure for clustering. The score is easily interpretable and rewards well separated clusters no matter the

[5] Recreated from https://scikit-learn.org/stable/modules/clustering.htmlsilhouette-coefficient

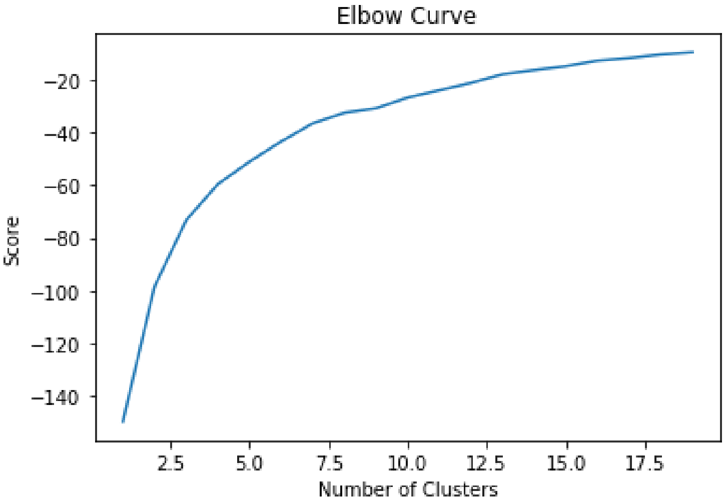


Figure 7: Attempting to use the elbow heuristic with k-means model selection.

amount of clusters. The silhouette score indicates what k could be the best clustering.

No indication of a good silhouette score was found and if we had found one it would likely be highly problem-specific. Instead, we will create our own indication by artificially creating a favourable scenario for our model. In this scenario a patient who scores the highest in one feature, scores the highest in all features. The second-highest rank would score the second-highest in all other features, and so on. A clustering algorithm would then easily divide the subjects with lower feature scores from the subjects with higher feature scores. The average silhouette score of this favourable scenario will be calculated as a guideline for a good model. This model will be called the strictly ordered model and its silhouette score will be compared to the real scenario.

3.2.2 Hierarchical Clustering validation

For HC we will use an intuitive rule of thumb and a normalized validation measure, as well. The added value of hierarchical grouping is the way the dendrogram shows the creation of the clusters. The dendrogram shows the relative distance between data points and pairs data points based on this. On the x-axis the data points are all singletons and they are grouped along the y-axis in a tree structure. It provides a clear overview of the memberships that were formed, as well as a rule of thumb for the best clustering. Cutting it at the right place will result in the right partition of the data. Given a dendrogram the cut should be made where the vertical line can be longest without branching off. The amount of clusters is then the amount of vertical lines that is crossed by an imaginary line (in red in figure 8[6]). In figure 9 it is visible why the choice is not always obvious. The dendrogram cut-off seems to be as arbitrary as the elbow method.

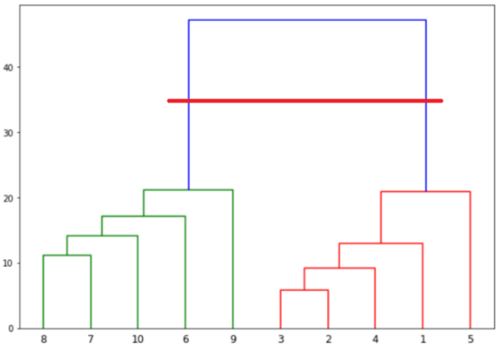


Figure 8: Cutting a dendogram for model selection.

[6] Example dendrogram https://stackabuse.com/hierarchical-clustering-with-python-and-scikit-learn/

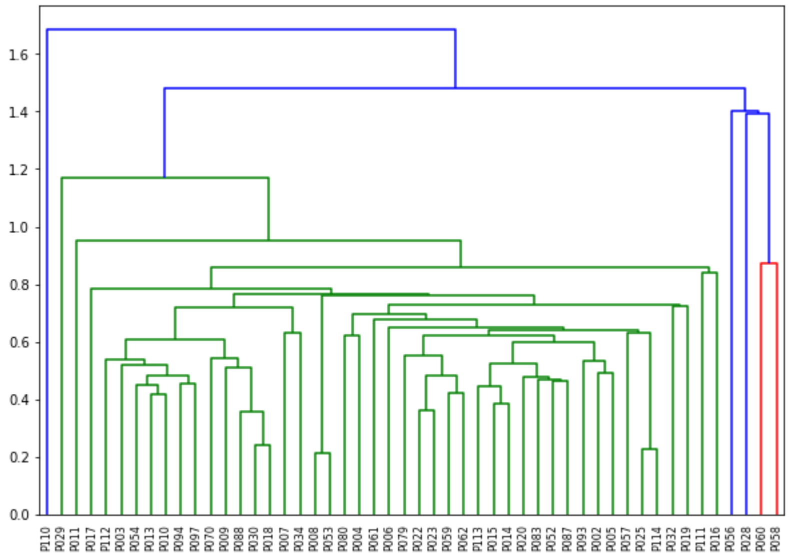


Figure 9: Dendrogram produced with all features. Where should the cut be made?

Therefore, the normalized validation measure cophenetic correlation coefficient is used for HC validation. It calculates the correlation between the original absolute distances between points and the cophenetic distance. The cophenetic distance between two points is the vertical distance in the dendrogram before they are merged together bottom-up. It evaluates how well these dendrogram distances represent the absolute distances. If the distances correlate well, the hierarchy of choices represents the closeness in absolute distance well. The cophenetic correlation coefficient c is defined as:

$$Cophenetic\ correlation\ coefficient\ =\ \frac{\sum_{i<j}(Y_{ij}\ -\ y)(Z_{ij}\ -\ z)}{\sqrt{\sum_{i<j}(Y_{ij}\ -\ y)^{\mathbf{2}}\sum_{i<j}(Z_{ij}\ -\ z)^{\mathbf{2}}}}$$

Y$_{ij}$: Euclidean distance between objects i and j

Z_{ij} : Cophenetic distance between points i and j

y : average of Y

z : average of Z

This formula[7] provides the coefficient of correlation between the set Y of Euclidean distances between points and the set Z of Cophenetic distances between points. A Cophenetic coefficient of almost 1 indicates a high quality cluster, because the real distance is closely reproduced by the dendrogramatic distance. This means that the cluster choices based on the cophenetic distances will also work well for the original points. It is said that c = 0.75 is a guideline for a good clustering model (Hintze, 1992). The cophenetic correlation coefficient will help us decide what cluster model performs well, and the dendrogram will decide how many clusters will be produced. If a cluster is very small, it is added to the closest cluster, so the membership comparisons are still meaningful.

3.2.3 Membership similarity

A measure for comparing memberships between clusters is useful for model selection in two ways. If the clusters that two models produce have very similar members, their models may be using features with similar information. This information may be combined to distinguish the clusters even better. If two models both produce clusters which distinguish similar diagnostic or demographic properties, but they have little membership similarity, the features could be combined to magnify this effect. We introduce the highest cluster similarity score to calculate the similarity of the most matching

[7] Formula rewritten from https://nl.mathworks.com/help/stats/cophenet.html

clusters from two models between zero and one. Effectively, this means that the most similar cluster pair is evaluated between two models. We introduce c as the list of symmetric set differences between any one cluster from the original model C_1 and any one cluster from the compared model C_2 for every combination. c is the list of $length(c_1 \oplus c_2)$ for every pair in $c_1 \in C_1 \times c_2 \in C_2$. The minimum difference c_{top} in members between the two most similar clusters is then:

$c_{top} = min(c)$.

We then introduce the highest cluster similarity score for a model with N=50 data points as: $\mathbf{1} - \frac{c_{top}}{N}$. For example, if two cluster models have one equal cluster (with equal members), their smallest set difference c_{top} will be 0, and their highest cluster similarity is $1.0 - \frac{0}{50} = 1.0$.

3.2.4 Solution quality

Solution quality, or model quality, is a subjective measure which will be used in addition to the other model evaluation measures. The solution quality will be determined according to our goal, because if a cluster model has a high validation score, but does not produce interesting results, it does not have to be the best model for our purpose. Therefore, the resulting clusters will be compared in terms of diagnostic scores. A model which shows interesting diagnostic differences, without confounding demographics will have good solution quality. If it then also has good validation scores it is a good candidate to become the best model. Solution quality is determined by how well a model can distinguish members with different diagnostic properties. If a model separates members that have a high chance of being from a different distribution of PANSS scores, then that matches with our goal. If a model finds two clusters the non-parametric Mann-Whitney U test for significance will be used to analyze property differences. For each cluster the compared property is grouped. The U test then tests if a property's mean for both clusters μ_1 and μ_2 are the same: H0: $\mu_1 = \mu_2$. H1: $\mu_1 \neq \mu_2$. The significance level is .05. If more clusters are found, the ANOVA test of significance will be applied to test for differences in distributions.

3.3 Model selection

We will then perform model selection based on model evaluation. The models will be created from the selected features. The search for the best combinations of features can be performed top-down and bottom-up. The best choice is to start with one feature models and build from there to develop a feel for the individual features, because the research is explorative. Even though correlation measures might better describe the effects of one-dimensional features than clustering, clustering will be used so model exploration can be done with a consistent method. This allows for easy and neat comparison, as well as membership comparison. One-dimensional features will only be used to explore the effects of the features and will not be considered as best models. The aim is to explore the multivariate effects. Combinations of well-performing features can then be made. Naturally, not all combinations of features make sense. The *variance* and *standard* deviation will not be combined, just like the *median* and *mean*. These measures will be used separately for performance comparison.

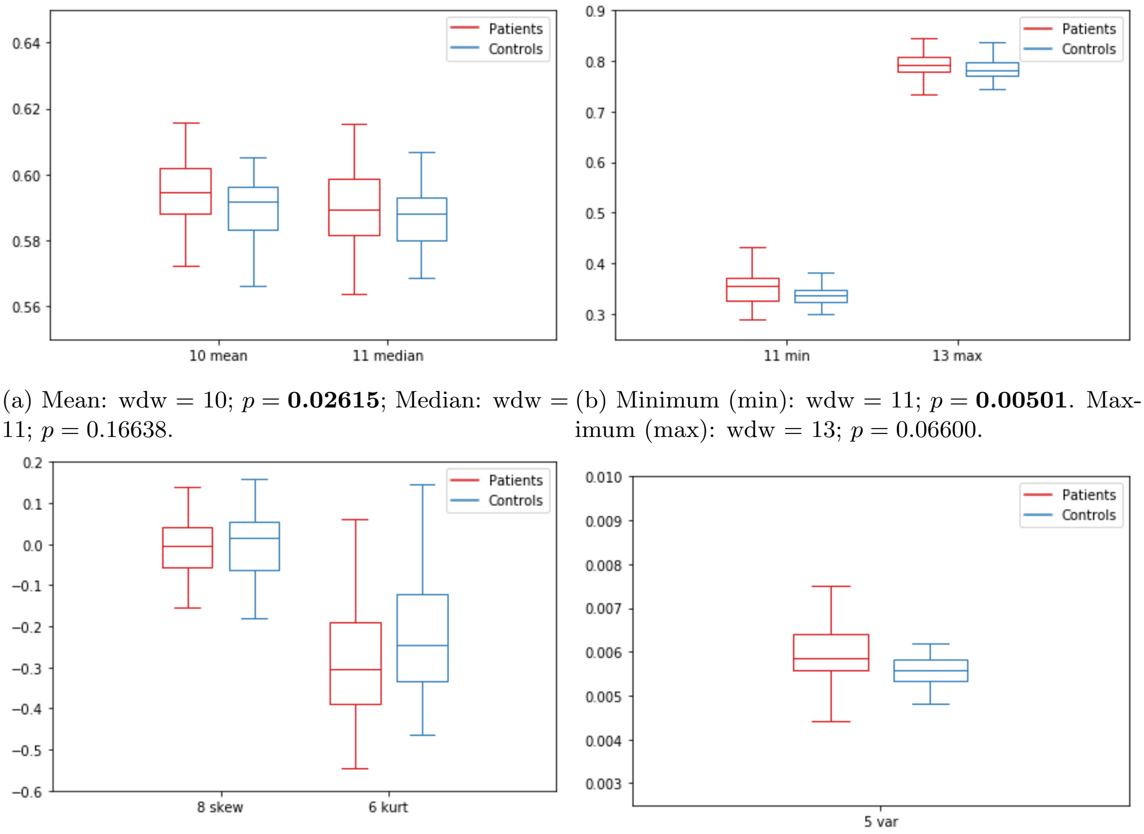
The model selection will be guided by model evaluation. Model evaluation will compare models in terms of memberships, cluster validation scores and solution quality. Uninformative or misinformative features will be removed along the way. All data will be standardized and shuffled before clustering. First, the selected features with the highest solution quality will be chosen. From there, using model selection, one top model will be chosen per cluster algorithm based on cluster validation score and solution quality. The final top model(s) will be visualized. If more than two features are used, PCA can reduce this to two dimensions for visualization. Ultimately, the diagnostic, demographic and linguistic differences between clusters will be discussed.

4 Results

4.1 Feature selection

The feature selection results in one window size (wdw) per feature. The test of significance test between patients and controls is shown in p-values.

Many features report low p-values in the test of difference between patients and controls. *variance*, *minimum*, *kurtosis* and *mean* show to be significantly different between groups. These features



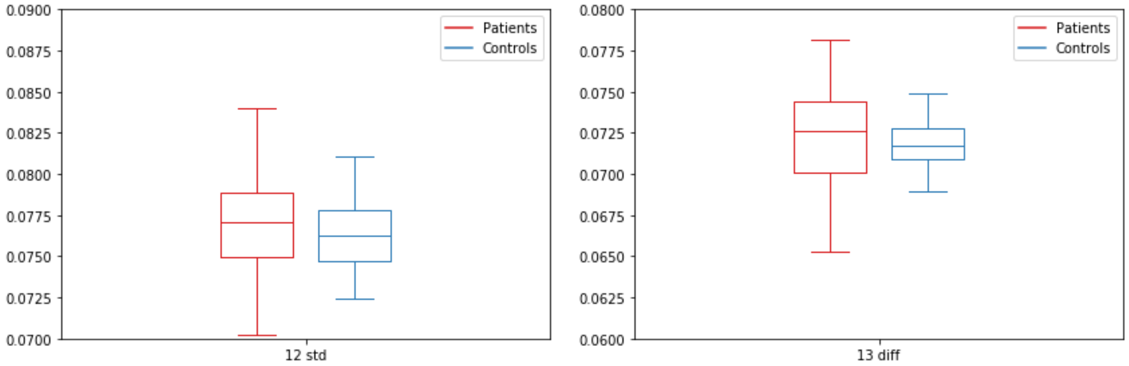
(a) Mean: wdw = 10; $p =$ **0.02615**; Median: wdw = 11; $p = 0.16638$.

(b) Minimum (min): wdw = 11; $p =$ **0.00501**. Maximum (max): wdw = 13; $p = 0.06600$.

(c) skewness (skew): wdw = 8 ; p = .21295. kurtosis (kurt): wdw = 6; $p =$ **.01790**.

(d) Variance (var): wdw = 5; $p =$ **.0003**.

are thus validated. For all features, but the time series features, the median of all values is higher for patients. An interesting trend is that the box plot distributions of patient data are more spread out.



(a) Standard deviation (std): wdw = 12; $p = .06778$.

(b) Difference (diff): wdw = 13; $p = .16296$.

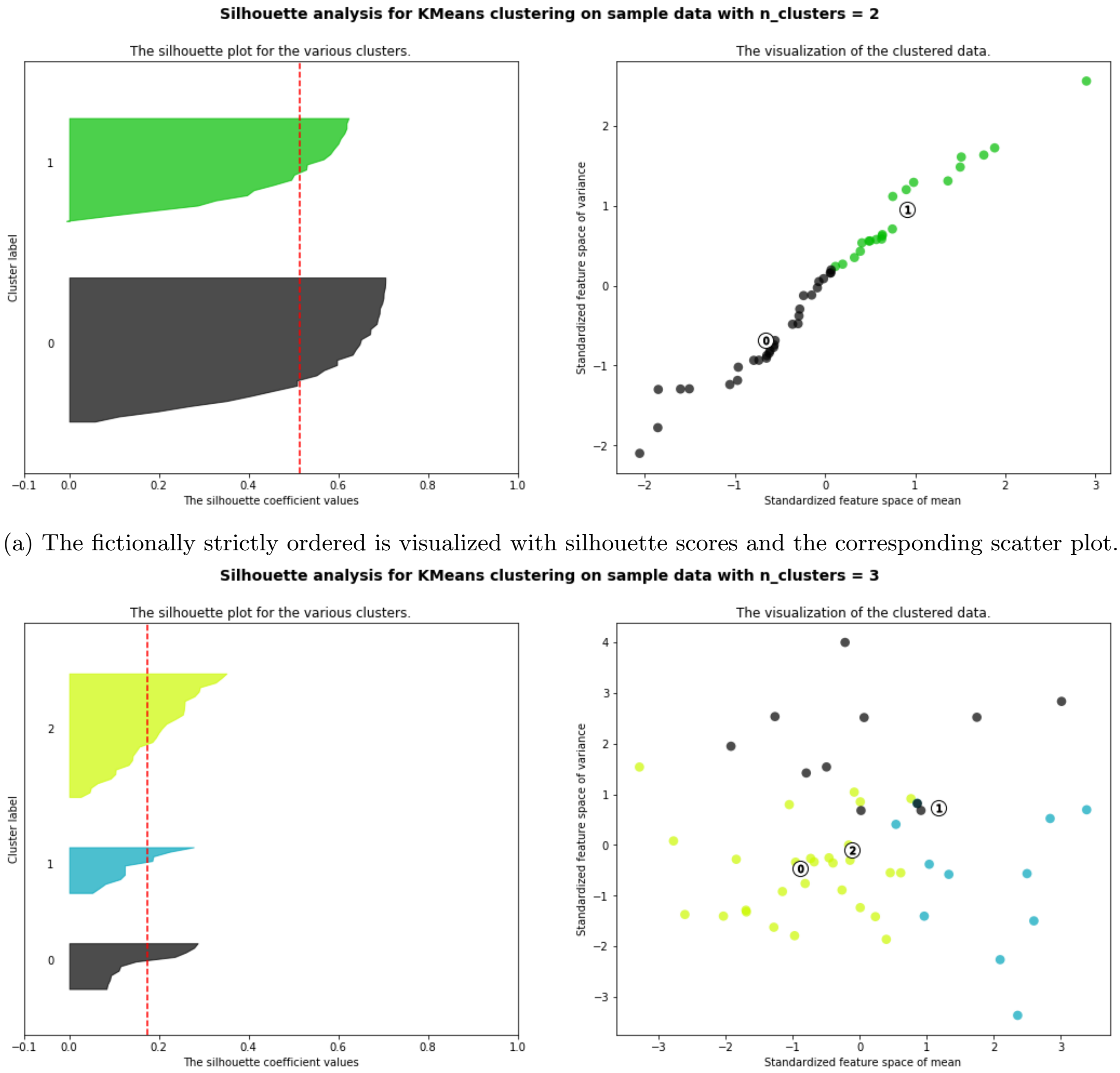
Figure 11: The distributions of each feature with the selected window size compared between patients and controls.

4.2 Clustering results

4.2.1 K-means clustering exploration

The favourable scenario is created and cluster validation is performed. This is compared to a model using all features in original order (12a, 12b. The latter is not performing very well with an average silhouette score of only .19 versus .52. The cluster average score is closer to 0 than to the strictly ordered set with the full feature set. $s = .52$ will be used as a rule of thumb. All k-means models used random restart with 500 restarts, a value at which the validation scores seemed to converge. This value was found empirically.

The basis of our model selection will be chosen from the boxplot results. *mean*, *variance*, *minimum*, *kurtosis* are significantly different between patients and controls (10a, 10d, 10b, 10c) and will form the fundaments for bottom-up search model selection. K-means is performed on these features individually. The models of *variance*, *minimum* and *kurtosis* do best at $k = 2$, while mean is clustered at $k = 3$. We use our three model evaluation criteria. The cluster validation scores are good according to our rule of thumb, but do not differ too much (between $s = .5$ and $s = .6$ (A.3)). The highest cluster similarities between the *mean* cluster model and the *minimum*, *variance*, and *kurtosis* models are well above the median highest cluster similarity of 0.58 (A.1). In terms of solution quality all four singular feature cluster models show a visible inter-cluster PANSS score difference, except for *kurtosis* (Figure 13a, 13b, 13d, 13e). Besides, clustering on $difference$ also shows a discrepancy in PANSS general (Figure 13c). The similar diagnostic properties and relatively high highest cluster similarities indicate the combination of these features could improve the model.

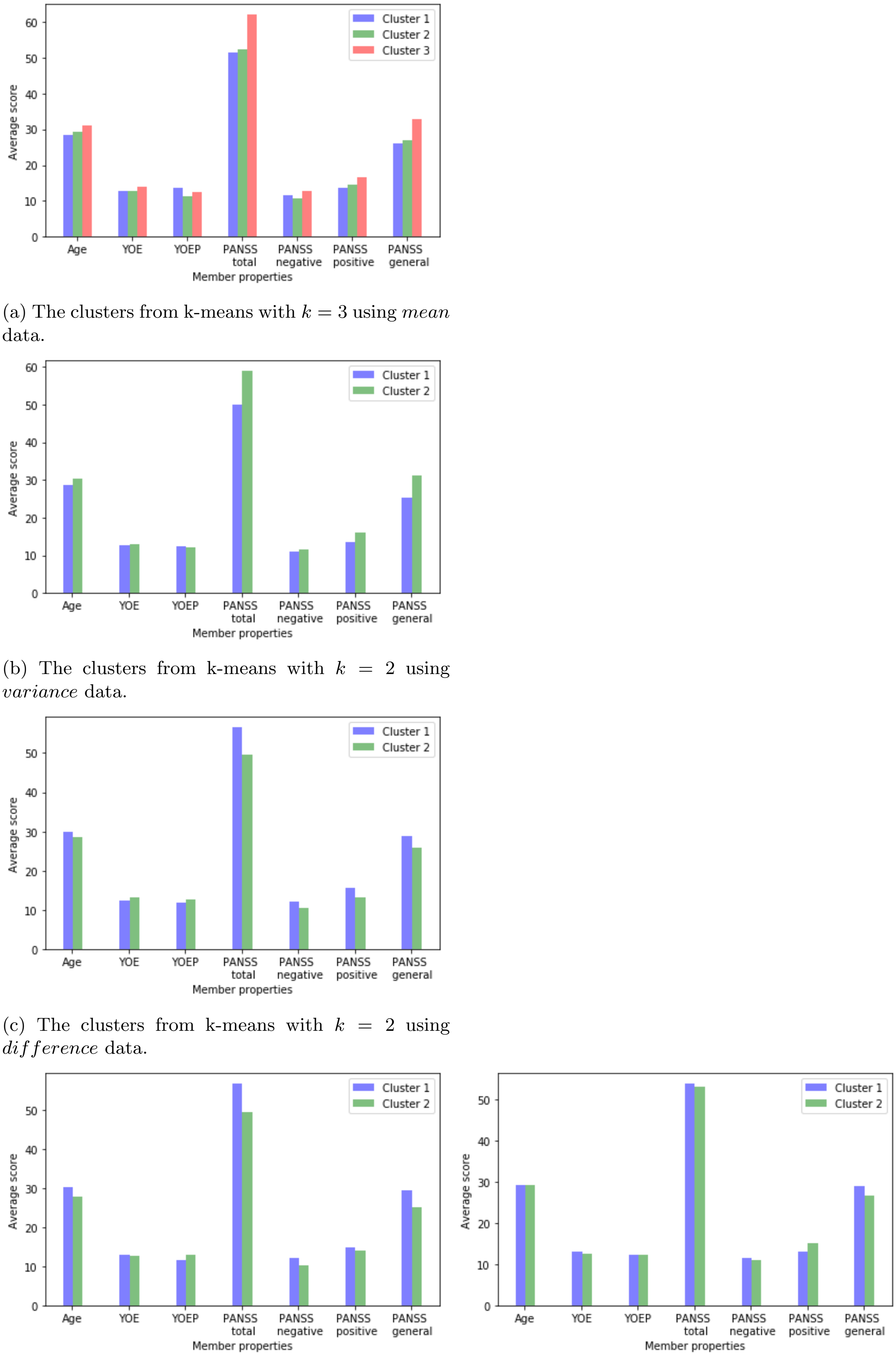


(a) The fictionally strictly ordered is visualized with silhouette scores and the corresponding scatter plot.

(b) The real data is shown with silhouette scores and the corresponding scatter plot.

Figure 12: This figure shows the comparison of k-means modelling on a data set with patients that have a strict coherence ranking versus the real data.

Then, we combined the features in k-means. The highest cluster similarity values between the one-dimensional models are relatively high, but the absolute values are low. So, the silhouette coefficient is likely to suffer, because they cluster members differently. In indeed all cases adding more features dramatically lowers the score as seen in the full table (A.3). However, perhaps these models distinguish between groups with different diagnostic scores better. Therefore, one model with the highest silhouette score of 2, 3 and 4 features will be picked. The cluster quality of the resulting models will be assessed to pick a top model for k-means.



(a) The clusters from k-means with $k = 3$ using *mean* data.

(b) The clusters from k-means with $k = 2$ using *variance* data.

(c) The clusters from k-means with $k = 2$ using *difference* data.

(d) The clusters from k-means with $k = 2$ using *minimum* data.

(e) The clusters from k-means with $k = 2$ using *kurtosis* data.

Figure 13: k was based on the best silhouette coefficient and k-means was performed per individual feature. The average demographic and diagnostic properties of members per cluster are shown.

4.2.2 Hierarchical Clustering exploration

The bottom-up search for HC model selection had similar results to k-means model selection. We again apply our three measures of model evaluation. We again found the cluster validation measures (cophenetic scores in this case) are high for models using *mean*, *difference*, *variance* or *minimum*. All individually show discrepancies in PANSS scores between clusters. *variance* results in the highest PANSS general score difference, similarly to k-means. The highest cluster similarity between these four models are also very similar to those of the k-means models (A.2).

After combining features the cophenetic score unexpectedly goes up to $c = .72$ for [*mean*, *variance*] or even $c = .81$ for [*mean*, *difference*, *variance*, *minimum*] (Figure 14 and 15). The model with the highest validation score again uses *variance* and *mean*. Perhaps, the models have found similar solutions to the problem. The solution quality of the best model for 2, 3 and 4 features will be analyzed further by solution quality.

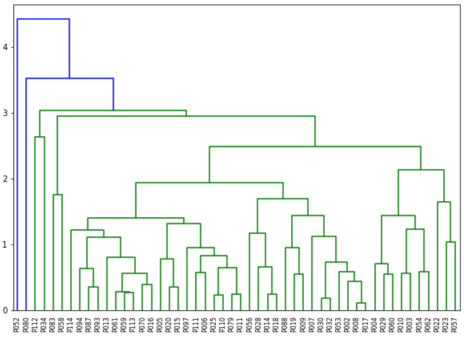


Figure 14: Dendrogram using features mean and variance: cophenetic score .81.

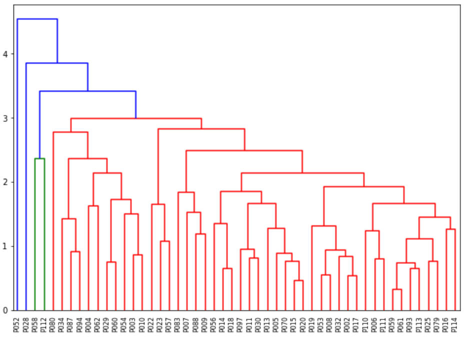


Figure 15: Dendrogram using features mean, difference, variance and minimum: cophenetic score .72.

4.3 Model selection with solution quality

4.3.1 K-means clustering properties

Three models with varying performances have resulted from the k-means model exploration. Model quality and validation scores will decide what model performed best. The cluster model on features [*mean*, *difference*, *variance*, *minimum*] produced two clusters with a significantly different PANSS general score, and thus has a good chance of separating members with different PANSS scores. However, the silhouette score was only .251. Both the significant effect and the silhouette score is trumped by the model using [*mean*, *variance*]. This model is therefore the best model of k-means model selection. The full table of means and standard distributions can be found in appendix 2.

K-means clustering with $N = 50$ on features [*mean*, *variance*] with the best validation score for $k = 2$ has resulted in a significantly different mean PANSS general score between clusters ($\alpha = .05$) (table 2) with a difference in mean of $||\mu_1 - \mu_2|| = 6.90$ and an effect size Cohens $d = 0.9337$. This is considered a strong effect (Sawilowsky 2009).

Table 2: K-means model cluster comparison for the best model of two, three and four features.

	Silhouette score (s)	PANSS total score	PANSS positive score	PANSS negative score	PANSS general score
mean & variance	.490	.0651 (t=161.0)	.2946 (t=204.0)	.1447 (t=181.0)	**.0072** (t=120.0)
mean & difference & minimum	.316	.4494 (t=207.0)	.3610 (t=266.5)	.3511 (t=203.5)	.3550 (t=182.5)
mean & difference & variance & minimum	.251	.1838 (t=219.5)	.4111 (t=251.5)	.2518 (t=230.5)	**.0209** (t=166.0)

The three k-means models were tested for confounding factors. For all models a significantly different ratio of men to women was found. There were only 8 women in the data set, however (table 3).

Table 3: K-means cluster models demographic features comparison.

	Gender	Age	YOE	YOEP
mean & variance	**.0001** (t=56.4)	.1725 (t=186.0)	.4492 (t=222.0)	.4909 (t=226.5)
mean & difference & minimum	**<.0001** (t=15.1)	.4714 (t=276.5)	.3689 (t=264.0)	.2028 (t=263.5)
mean & difference & variance & minimum	**.0002** (t=44.4)	.3319 (t=241.5)	.2245 (t=227.0)	.1492 (t=211.0)

4.3.2 Hierarchical Clustering clustering properties

The three best models that have resulted from HC exploration will be compared in terms of solution quality and validation scores. The results of solution quality assesment are showed as the significance test results of diagnostic score differences between clusters (table 4). The model using features [*mean*, *difference*, *variance*, and *minimum*] was chosen, because it showed the strongest evidence that the members of different clusters had different PANSS scores (general PANSS scores). The cluster validation score was also close to the mentioned rule of thumb $c = .75$.

Table 4: HC best cluster models for two, three and four features.

	cophenetic score (c)	PANSS total score	PANSS positive score	PANSS negative score	PANSS general score
mean & variance	.81	.1623 (t=172.0)	0.5 (t=214.0)	.2478 (t=185.0)	**.0258** (t=131.0)
mean & variance & minimum	.76	.1100 (t=149.0)	.4415 (t=193.5)	.2796 (t=175.5)	**.0139** (t=109.0)
mean & difference & variance & minimum	.72	.0524 (t=185.5)	.1217 (t=207.5)	.2351 (t=228.0)	**.0074** (t=147.0)

In table 5 the demographic features of the HC models are compared between clusters. In the clusters that resulted from Hierarchical Clustering again no significant effect between members was found other than the gender category. Because the count of women in this category was too low, it is again not good evidence of a relationship between gender and coherence.

Table 5: Hierarchical Clustering cluster models demographic features comparison.

	Gender	Age	YOE	YOEP
mean & variance	**<.0001**	.3029	.4666	.3981
	(t=72.5)	(t=192.0)	(t=210.5)	(t=203.0)
mean & variance & minimum	**<.0001**	.1426	.4070	.4420
	(t=92.5)	(t=155.5)	(t=190.0)	(t=193.5)
mean & difference & variance & minimum	**<.0001**	.2354	.2943	.2833
	(t=257.0)	(t=228.0)	(t=237.0)	(t=235.0)

Hierarchical Clustering was performed on the set of 50 patients using [$mean$, $difference$, $variance$, and $minimum$] with a cut-off at two clusters. The members of the two clusters were subjected to a significance test to test for PANSS score differences. This resulted in a significant difference in PANSS general score ($\alpha < 0.05$). The difference in mean PANSS general score was $||\mu_1 - \mu_2|| = 6.26$ and effect size Cohens $d = 0.8685$ between members of the clusters. This is considered a strong effect (Sawilowsky 2009).

4.3.3 Comparison of best models

Five models found clusters with a significant difference in PANSS general score. From each algorithm the model was chosen that performed the best according to our predefined model evaluation criteria. Two models that were most likely to separate clusters with different diagnostic properties and had good cluster validation scores were chosen as top models. K-means selected a model using two features, while HC selected a model using four features. The most similar cluster pair between these two models differed 9 data points, meaning the highest cluster similarity between these two models is .82. These two similar clusters will be called the Cluster 0 of the model for comparison.

The centroid coordinates are represented in table 6 for k-means and table 7 for HC. The higher the value of a centroid's dimension, the more the feature contributed to the clustering. The larger the difference of a feature between clusters, the better the clustering is separated in this dimension. The algorithm for HC does not provide a centroid as k-means does. For fair comparison the centroid was calculated for the clusters and formatted in the same way. For k-means members of Cluster 0 are characterized to have higher *mean* and higher *variance*, as well as a higher PANSS general score (mean absolute difference 6.90). For HC the members of Cluster 0 have higher *mean*, $difference$, *variance* and *minimum*, as well as a higher average PANSS general score than Cluster 1 members (mean absolute difference 6.26).

Table 6: K-means clustering top model centroid coordinates in standardized feature space.

	Cluster 0	Cluster 1
Mean	1.2676	-0.4003
Variance	1.2063	-0.3809

Table 7: Hierarchical Clustering top model calculated centroid coordinates in standardized feature space.

	Cluster 0	Cluster 1
Mean	0.6449	-0.2764
Difference	0.4830	-0.2070
Variance	0.7301	-0.3129
Minimum	0.5191	-0.2225

The k-means and HC top model are visualized. K-means is visualized using the 2D space the algorithm already used: *mean* and *variance* (Figure 16). The HC model uses four features, however. PCA could not be used to visualize the HC model in 2D, because only 65% variance was explained by two principal components (A.12). Therefore, the 4 dimensions were split up in six 2D visualizations (Figure 17a through 17f).

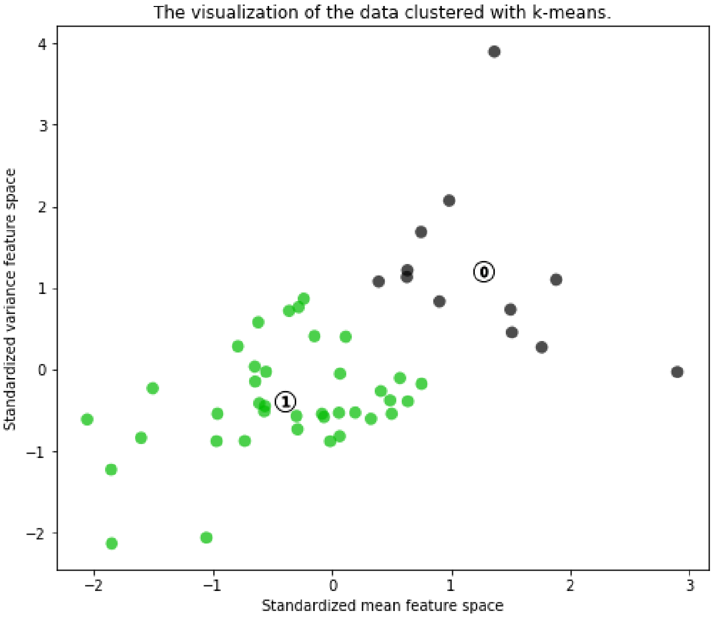
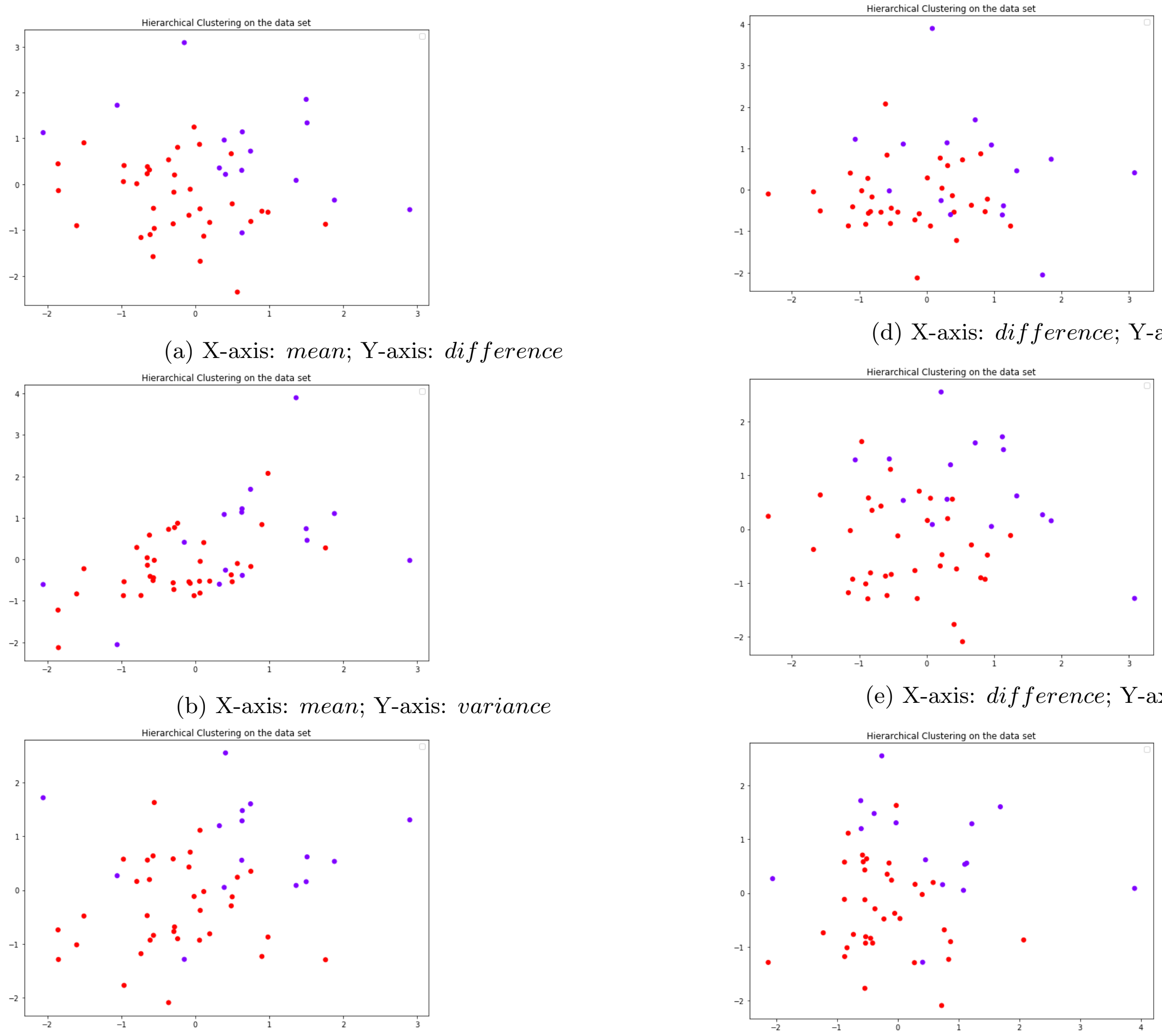


Figure 16: K-means best model using mean and variance visualized in 2D.



(a) X-axis: $mean$; Y-axis: $difference$

(b) X-axis: $mean$; Y-axis: $variance$

(c) X-axis: $mean$; Y-axis: $minimum$

(d) X-axis: $difference$; Y-axis: $variance$

(e) X-axis: $difference$; Y-axis: $minimum$

(f) X-axis: $variance$; Y-axis: $minimum$

Figure 17: Hierarchical Clustering best model used mean, difference, variance, and minimum. The 4D feature space was reduced to six 2D visualizations.

4.3.4 Confounding factors

No significant differences on group level between clusters were found among the demographic features, so no evidence was found that the included demographic features are confounding factors. Although we did not find any evidence that significant effects were confounded by demographics, this does not exclude that there could be multivariate effects confounding the results. Because there was no reason medicine D2R categories could be confounding results, these were not included in the results (A.10, A.11).

5 Discussion

We explored the relationship between linguistic coherence and diagnostic scores with unsupervised learning for subjects with psychotic disorder. First, word2vec semantic space modelling was used to convert the text from interviews to vectors. Then, the vectors were grouped into windows to split up the interview into smaller parts. Afterwards, the cosine similarity was calculated for every window of vectors. These cosine similarities formed the raw coherence data features were extracted from. Nine general purpose and time series features were extracted and optimized to separate patients from controls. These features aimed measure language incoherence. Unsupervised cluster models were used to distinguish different groups of patients based on language coherence.

For every feature a window size was found to maximize the difference between patients and controls. Four features were found to be significantly different between patients and controls. These features could potentially represent different elements of linguistic coherence. Model selection with these features among others and the clustering algorithms resulted in models creating two clusters from the patients measured. Interestingly enough, the two solutions were similar, but not the same. Both were validated by cluster validation measures. This means that the patients could be divided into two groups with different language aspects. Not only did these two groups have different language features, the groups were also found to have a significantly different PANSS diagnostic score with a strong effect size. This implies that patients in this data set with different language coherence properties have different symptom scores. Consequently, this is evidence that there is a relation between linguistic coherence and diagnostic scores.

Ultimately, these language properties were analyzed and patients with higher general symptom scores turned out to have more variance in their linguistic coherence. They also spoke more coherently on average. One model suggested they also had a coherence level that went up and down more and that the least coherent window of words was relatively coherent opposed to people with a lower general symptom score. We speculate this means that patients suffering more from general symptoms have more turbulent speech. This language could vary between very coherently and very incoherently, which would also explain the higher $difference$. Perhaps the subjects speak more coherently on average, because they experience preoccupation, a general symptom, causing them to obsess with the specific subject at hand. It is counter-intuitive that the minimum cosine similarity of the windows is higher with higher general symptom scores. However, as mentioned the $minimum$ measure is very prone to outliers and because of that not a statistically reliable measure. These results suggest it should either be used with a larger window size to decrease the weight of the outlier or not at all.

This is proof of concept that two groups with different linguistic coherence properties may exist within patients with psychotic disorder. Moreover, we found that these two groups also experience different severity in general symptoms. A possible explanation is that some general symptoms, such as unusual thought content, preoccupation, disorientation and poor attention span could cause abnormal language patterns. No evidence for confounding demographic features was found. This suggests that linguistic coherence and general symptoms score are related. Although the results did not lead to hard evidence, the aim of this research was exploration. These findings are the first step to using language as biomarker for detection of illness severity. This automated data-driven approach could increase timely care for sufferers from psychotic disorder.

6 Limitations and suggestions for further research

Topic transition could be a problem in interview analysis. Every collected interview was different in size and contained multiple different questions. A complex problem is the overlap between two different questions, because of the topic switch the answers will also contain very different words. The sliding window might contain two answers about two different topics, resulting in low coherence scores. It could also be that patients are less talkative and require more questions per interview. This would cause more of these noisy incoherence windows, which do not represent actual linguistic incoherence. This is a problem for validating the outlier sensitive measures maximum and minimum coherence. These features are guaranteed to have extreme values due to topic transition. Another problem are stop phrases, such as "oh oh", that result in high maximum cosine similarities for many low window sizes.

Cosine similarity was used to produce one value out of the vectors in a window of word vectors, but some argue cosine similarity has fundamental disadvantages, causing it to be biased towards longer sentences and non-informative utterances, assigning higher coherences to these (Iter, Yoon & Jurafsky, 2018). This problem of 'verbal fillers' is recognized and was dealt with by using larger window sizes. It does not seem to be a fundamental problem with simple preprocessing as mentioned in the paper. The best distance measure should be explored more thoroughly for more accurate incoherence representation, but in our explorations we found larger coherence windows were not assigned higher coherence scores. The paper does not elaborate on the reasoning behind this, however.

It should also be noted the PANSS score, as well as the interview itself, is a snapshot of a patient's status. Psychotic disorder is a syndrome which is examined over time, while the PANSS score and the interview are assessed at one moment. It should not be mistaken for a direct degree of psychotic disorder. Language is powerful in that it is easily obtainable and could have predictive powers for psychotic disorder onset, but there is no explicitly proven relationship between PANSS scores and linguistic coherence.

Despite the advantages of clustering, the subjectivity of cluster interpretation and cluster quality is a problem, which ironically invokes some bias. Although rules of thumb were helpful, it should be noted these are just proposed heuristics. The k-means algorithm with random restart was not stable, requiring as many as 500 iterations to provide reliable scores. This could indicate the data is hard to cluster. Interestingly enough, the models clustering on two features resulted in the highest validation scores.

Many tests were executed in the study for the sake of exploration. This is problematic, because with many trials some rule of thumb will always be fulfilled by pure chance. The data set was relatively large, compared to previous studies using between 5 and 24 samples [Iter et al., 2018]. Despite the explorative nature of the study and multitude of statistical tests, the significance level was set at .05 to allow more results to be explored further.

The demographic data was analyzed to check for confounding factors. For the male to female ratio this is hard to say, because of the imbalance in the data set. No evidence for other confounding factors or demographic differences was found in the interpreted clusters, but there could be a multivariate effect of demographic features on spoken language coherence. For example, age and years of education together could confound coherence differences. In future studies the effect of demographics on coherence should be extensively analyzed to ensure the legitimacy of the results.

Medicine usage was not included in the results due to the complexity of neuropharmacy and the lack of evidence that it could be confounding linguistic incoherence. The diagnostic measures used are symptom-based scores and medicine are partially prescribed exactly to counter these symptoms. Certain symptoms might cause certain medicine to be prescribed and these medicine might impact the symptoms again. This complicated causal chain makes it difficult to account for medicine usage. However, as stated research suggests that medicine usage could possibly confound language semantically. During data analysis it was found that models no longer found a significant effect when split in two categories of high and not high binding of the type of antipsychotic medicine to the dopamine D_2 receptor (A.11, A.10).

The feature set was restricted to one window size per feature, although the mean coherence of four words could contain very different information from the mean coherence of 15 words. One gives a measure of higher level context than the other. In particularly for the successful features multiple window sizes could be combined in future research to provide a more sophisticated representation of the feature. With more informative features future endeavours could successfully define the relationship between diagnostic scores and coherence measures. This could make detection of symptoms or amount of symptoms possible, paving the way for timely care without the need of expert-led interviews.

Acknowledgement: I would like to kindly thank H.G. Schnack for his extensive supervision, both H.G. Schnack and T. van Ommen for their feedback and A. Voppel from Universitair Medisch Centrum Utrecht for his insights on the data collection, the provision of the data set and generously sharing his domain expertise with me.

Conflict of interest: the author declares no conflict of interest.

7 Conclusion

In this research we discovered four features that could have the potential to represent coherence in spoken language. We found results suggesting patients with psychotic disorder can be divided into multiple categories based on language coherence properties. The results suggest a relationship between the coherence of spoken language and the diagnostic scores for psychotic disorder using unsupervised clustering. The features that contributed to this difference in diagnostic scores were identified. Nevertheless, we should be very careful to interpret these results, because of the explorative nature of this study. These findings make a plausible case that diagnostic scores and language coherence are related. In future research it can be investigated how exactly and why diagnostic scores and language coherence could be related. This could form the basis for early detection of worsening psychotic symptoms without the need of intensive expert-lead interviews.

References

[Adam, 2013] Adam, D. (2013). Mental health: On the spectrum.

[Allen, 2019] Allen, T. A. (2019). Information technology laboratory.

[Altszyler et al., 2016] Altszyler, E., Sigman, M., Ribeiro, S., and Slezak, D. F. (2016). Comparative study of lsa vs word2vec embeddings in small corpora: a case study in dreams database. *arXiv preprint arXiv:1610.01520*.

[Bedi et al., 2015] Bedi, G., Carrillo, F., Cecchi, G. A., Slezak, D. F., Sigman, M., Mota, N. B., Ribeiro, S., Javitt, D. C., Copelli, M., and Corcoran, C. M. (2015). Automated analysis of free speech predicts psychosis onset in high-risk youths. *npj Schizophrenia*, 1:15030.

[Bedi et al., 2014] Bedi, G., Cecchi, G. A., Slezak, D. F., Carrillo, F., Sigman, M., and De Wit, H. (2014). A window into the intoxicated mind? speech as an index of psychoactive drug effects. *Neuropsychopharmacology*, 39(10):2340.

[Brodersen et al., 2014] Brodersen, K. H., Deserno, L., Schlagenhauf, F., Lin, Z., Penny, W. D., Buhmann, J. M., and Stephan, K. E. (2014). Dissecting psychiatric spectrum disorders by generative embedding. *NeuroImage: Clinical*, 4:98–111.

[Corcoran et al., 2018] Corcoran, C. M., Carrillo, F., Fernández-Slezak, D., Bedi, G., Klim, C., Javitt, D. C., Bearden, C. E., and Cecchi, G. A. (2018). Prediction of psychosis across protocols and risk cohorts using automated language analysis. *World Psychiatry*, 17(1):67–75.

[de Boer et al., 2018] de Boer, J., Voppel, A., Begemann, M., Schnack, H., Wijnen, F., and Sommer, I. (2018). Clinical use of semantic space models in psychiatry and neurology: A systematic review and meta-analysis. *Neuroscience & Biobehavioral Reviews*, 93:85–92.

[de Boer et al., 2019] de Boer, J., Voppel, A., Wijnen, F., and Sommer, I. (2019). Language and speech disturbances in schizophrenia spectrum patients: Symptom or side-effect? *Manuscript submitted for publication.*

[Delespaul et al., 2013] Delespaul, P., de Haan, L., van Hoof, F., van der Gaag, M., Keet, R., Kroon, H., Mulder, N., van Os, J., Slooff, C., Sytema, S., van Weeghel, J., and Wiersma, D. (2013). Feiten en cijfers over psychose.

[Hintze, 1992] Hintze, J. L. (1992). *Number cruncher statistical system: statistical software*. Jerry L. Hintze.

[Hoffman et al., 2018] Hoffman, P., Loginova, E., and Russell, A. (2018). Poor coherence in older people's speech is explained by impaired semantic and executive processes. *eLife*, 7:e38907.

[Insel, 2010] Insel, T. R. (2010). Rethinking schizophrenia. *Nature*, 468(7321):187.

[Insel, 2014] Insel, T. R. (2014). The nimh research domain criteria (rdoc) project: precision medicine for psychiatry. *American Journal of Psychiatry*, 171(4):395–397.

[Insel, 2015] Insel, T. R. (2015). Raise-ing our expectations for first-episode psychosis.

[Iter et al., 2018] Iter, D., Yoon, J., and Jurafsky, D. (2018). Automatic detection of incoherent speech for diagnosing schizophrenia. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 136–146.

[Kay et al., 1987] Kay, S. R., Fiszbein, A., and Opler, L. A. (1987). The Positive and Negative Syndrome Scale (PANSS) for Schizophrenia. *Schizophrenia Bulletin*, 13(2):261–276.

[Laptev et al., 2015] Laptev, N., Amizadeh, S., and Flint, I. (2015). Generic and scalable framework for automated time-series anomaly detection. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1939–1947. ACM.

[Marquand et al., 2016] Marquand, A. F., Wolfers, T., Mennes, M., Buitelaar, J., and Beckmann, C. F. (2016). Beyond lumping and splitting: a review of computational approaches for stratifying psychiatric disorders. *Biological psychiatry: cognitive neuroscience and neuroimaging*, 1(5):433–447.

[Mikolov et al., 2013] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781.*

[Mota et al., 2012] Mota, N. B., Vasconcelos, N. A., Lemos, N., Pieretti, A. C., Kinouchi, O., Cecchi, G. A., Copelli, M., and Ribeiro, S. (2012). Speech graphs provide a quantitative measure of thought disorder in psychosis. *PloS one*, 7(4):e34928.

[Oostdijk, 2000] Oostdijk, N. (2000). Het corpus gesproken nederlands.

[Rezaii et al., 2019] Rezaii, N., Walker, E., and Wolff, P. (2019). A machine learning approach to predicting psychosis using semantic density and latent content analysis. *npj Schizophrenia*, 5(1).

[Sawilowsky, 2009] Sawilowsky, S. S. (2009). New effect size rules of thumb. *Journal of Modern Applied Statistical Methods*, 8(2):26.

[Schnack, 2017] Schnack, H. G. (2017). Improving individual predictions: machine learning approaches for detecting and attacking heterogeneity in schizophrenia (and other psychiatric diseases). *Schizophrenia research.*

[Skinner, 1981] Skinner, H. A. (1981). Toward the integration of classification theory and methods. *Journal of Abnormal Psychology*, 90(1):68.

[Sommer, 2017] Sommer, I. (2017). Psyfar. *Psyfar*, 4:68.

[Universitair Medisch Centrum Utrecht afd. psychiatrie,] Universitair Medisch Centrum Utrecht afd. psychiatrie. Cash questionnaire.

[Voleti et al., 2019] Voleti, R., Woolridge, S., Liss, J. M., Milanovic, M., Bowie, C. R., and Berisha, V. (2019). Objective assessment of social skills using automated language analysis for identification of schizophrenia and bipolar disorder. *arXiv preprint arXiv:1904.10622.*

[Voppel et al., 2019] Voppel, A., de Boer, J., Schnack, H., and Sommer, I. (2019). Computational semantics show robust incoherence in spectrum patients. *manuscript in preparation.*

[Wang and Krystal, 2014] Wang, X.-J. and Krystal, J. H. (2014). Computational psychiatry. *Neuron*, 84(3):638–654.

[Wu et al., 2008] Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Philip, S. Y., et al. (2008). Top 10 algorithms in data mining. *Knowledge and information systems*, 14(1):1–37.

[Yang et al., 2018] Yang, X., Macdonald, C., and Ounis, I. (2018). Using word embeddings in twitter election classification. *Information Retrieval Journal*, 21(2-3):183–207.

[Zeng et al., 2014] Zeng, L.-L., Shen, H., Liu, L., and Hu, D. (2014). Unsupervised classification of major depression using functional connectivity mri. *Human brain mapping*, 35(4):1630–1641.

A Appendix

A.1

Table A.1: K-means highest membership similarity per feature

	Mean	Med.	Diff.	Var.	Standard deviation	Min.	Max.	Skew.	Kurt.
Mean	1.0	0.54	0.52	0.68	0.54	0.62	0.6	0.58	0.64
Median	0.54	1.0	0.7	0.58	0.52	0.52	0.58	0.6	0.6
Difference	0.52	0.7	1.0	0.6	0.58	0.54	0.56	0.5	0.52
Variance	0.68	0.58	0.6	1.0	0.62	0.54	0.76	0.5	0.56
Standard dev.	0.54	0.52	0.58	0.62	1.0	0.6	0.66	0.52	0.66
Minimum	0.62	0.52	0.54	0.54	0.6	1.0	0.58	0.52	0.54
Maximum	0.6	0.58	0.56	0.76	0.66	0.58	1.0	0.54	0.72
Skewness	0.58	0.6	0.5	0.5	0.52	0.52	0.54	1.0	0.54
Kurtosis	0.64	0.5	0.52	0.56	0.66	0.54	0.72	0.54	1.0

A.2

Table A.2: Hierarchical Clustering highest membership similarity per feature

	Mean	Difference	Variance	Minimum
Mean	1.0	0.52	0.6	0.62
Difference	0.52	1.0	0.58	0.54
Variance	0.6	0.58	1.0	0.52
Minimum	0.62	0.54	0.52	1.0

A.3

Table A.3: K-means average silhouette score performance per high potential feature combination

mean	.53
difference	.57
variance	.61
minimum	.58
mean & difference	.33
mean & variance	.49
mean & minimum	.33
difference & variance	.37
difference & minimum	.32
variance & minimum	.35
difference & variance & minimum	.22
mean & difference & minimum	.25
mean & difference & variance	.32
mean & variance & minimum	.32
mean & difference & variance & minimum	.25

A.4

Table A.4: K-means model cluster comparison: *mean*, *variance* with $s = .490$.

	PANSS total score (mean)	PANSS positive score (mean)	PANSS negative score (mean)	PANSS general score (mean)
Cluster	50.8684 ($\sigma = 10.5260$)	11.0 ($\sigma = 3.993$)	14.1053 ($\sigma = 5.1185$)	25.7632 ($\sigma = 5.5981$)
Cluster	60.5833 ($\sigma = 15.0635$)	12.25 ($\sigma = 5.1660$)	15.6667 ($\sigma = 4.3461$)	32.6667 ($\sigma = 8.4097$)

A.5

Table A.5: K-means model cluster comparison: *mean*, *difference*, *minimum* with $s = .316$.

	PANSS total score (mean)	PANSS positive score (mean)	PANSS negative score (mean)	PANSS general score (mean)
Cluster 0	53.7143 ($\sigma = 13.9639$)	11.4286 ($\sigma = 4.1613$)	14.1429 ($\sigma = 4.4538$)	26.5 ($\sigma = 5.4252$)
Cluster 1	52.5455 ($\sigma = 10.2635$)	11.1364 ($\sigma = 4.5457$)	14.9091 ($\sigma = 5.5670$)	28.1429 ($\sigma = 8.0031$)

A.6

Table A.6: K-means model cluster comparison: *mean*, *difference*, *variance*, *minimum* with $s = .251$.

	PANSS total score (mean)	PANSS positive score (mean)	PANSS negative score (mean)	PANSS general score (mean)
Cluster 0	51.2 ($\sigma = 10.7538$)	11.1143 ($\sigma = 4.0972$)	14.2571 ($\sigma = 5.2172$)	25.8286 ($\sigma = 5.8235$)
Cluster 1	57.8666 ($\sigma = 14.7868$)	11.7333 ($\sigma = 4.8230$)	15.0 ($\sigma = 4.3665$)	31.1333 ($\sigma = 8.1311$)

A.7

Table A.7: HC model cluster comparison: *mean*, *variance* with $c = .81$.

	PANSS total score (mean)	PANSS positive score (mean)	PANSS negative score (mean)	PANSS general score (mean)
Cluster	51.8974 ($\sigma = 12.1735$)	11.3333 ($\sigma = 4.4453$)	14.2564 ($\sigma = 5.1376$)	26.3077 ($\sigma = 6.4655$)
Cluster	57.8181 ($\sigma = 12.4814$)	11.1818 ($\sigma = 3.9270$)	15.2727 ($\sigma = 4.3294$)	31.3636 ($\sigma = 7.5350$)

A.8

Table A.8: HC model cluster comparison: *mean*, *variance*, *minimum* with $c = .76$.

	PANSS total score (mean)	PANSS positive score (mean)	PANSS negative score (mean)	PANSS general score (mean)
Cluster 0	51.8 ($\sigma = 12.0358$)	11.275 ($\sigma = 4.4045$)	14.275 ($\sigma = 5.0744$)	26.25 ($\sigma = 6.3943$)
Cluster 1	58.8 ($\sigma = 12.6791$)	11.4 ($\sigma = 4.0546$)	15.3 ($\sigma = 4.5398$)	32.1 ($\sigma = 7.5160$)

A.9

Table A.9: HC model cluster comparison: *mean*, $difference$, *variance*, *minimum* with $c = .72$.

	PANSS total score (mean)	PANSS positive score (mean)	PANSS negative score (mean)	PANSS general score (mean)
Cluster 0	59.7333 ($\sigma = 15.1678$)	12.8667 ($\sigma = 5.6317$)	15.0667 ($\sigma = 3.8724$)	31.8 ($\sigma = 8.2720$)
Cluster 1	50.4 ($\sigma = 9.8972$)	10.6286 ($\sigma = 3.4317$)	14.2286 ($\sigma = 5.3777$)	25.5428 ($\sigma = 5.4424$)

A.10

Table A.10: Medicine group clustering results: Hierarchical Clustering (p-values)

	Cophenetic score (c)	PANSS total	PANSS positive	PANSS negative	PANSS general
High D2R	.79	.1591	.2821	.2835	.3091
Low D2R	.80	.1533	.6001	.3013	.2392

A.11

Table A.11: Medicine group clustering results: K-means (p-values)

	Silhouette score (s)	PANSS total	PANSS positive	PANSS negative	PANSS general
High D2R	.56	.0657	.4648	.0274	.0956
Low D2R	.34	.1724	.5	.0854	.1338

A.12

Table A.12: Hierarchical Clustering top model PCA analysis

	Mean	Difference	Variance	Minimum	Variance explained
Component 1	0.70	-0.05	0.64	0.32	.4051
Component 2	-0.04	0.96	0.21	-0.19	.2531
Component 3	-0.07	0.25	-0.34	0.90	.2401
Component 4	0.71	0.13	-0.65	-0.23	.1017