Aus der Klinik für Psychiatrie und Psychotherapie

der Universitätsmedizin der Johannes Gutenberg-Universität Mainz

Clustering and predicting antidepressant response in patients with major depressive disorder – a machine learning based secondary analysis of “The EMC trial”-data

Clusterbildung und Prädiktion von Therapieverläufen depressiver Patienten – Eine Sekundäranalyse der „The EMC trial“-Daten mit Methoden des Maschinenlernens

Inauguraldissertation

zur Erlangung des Doktorgrades der

Medizin

der Universitätsmedizin

der Johannes Gutenberg-Universität Mainz

Vorgelegt von

Friedrich Albert Christoph Duge

aus München

Mainz, 2020

Tag der Promotion: 08. Dezember 2020

Table of contents

Tables

Figures

Introduction

Major depressive disorder (MDD) is a common disease affecting the general population with a lifetime risk of approximately 16.2% (United States of America, 2003) [1] to 17.1% (Germany, 2004) [2] in developed countries. The prevalence of depressive disorders in 2015 was estimated to be 4.4% worldwide with the total number of people living with depressive disorders estimated to be 322 million [3]. There is evidence of the economic burden increasing in recent years with the total economic burden of individuals with MDD in the United States of America totalling $210.5 billion in 2010 [4]. Globally, depressive disorders are ranked as the single largest contributor to non-fatal health loss [3].

The large economic and public health statistics of depressive disorders are complemented by powerful statistics for individual patient outcome. In people who died by suicide in Mannitoba, Canada between 1995 and 2009, depression was significantly more common than in matched controls with an adjusted odds ratio of 3.9 (95% CI: 3.35-4.52) [5]. An analysis of German health insurance data between 1987 and 1996 showed a significantly increased risk for permanent disability of patients that got treated for depression in an outpatient or inpatient setting compared to controls, with relative risks being 1.77 (95% CI: 1.56-2.00) and 3.47 (2,34-4,59) respectively.

These examples of statistics on outcome and disease burden underline the need for effective treatment intervention. Current German treatment guidelines for unipolar depression recommend psychotherapy or pharmacological therapy for moderate symptom severity and a combination of both for severe symptoms [6, p. 61]. Pharmacological therapy commonly involves antidepressants as first-line therapy [7], a substance class for which efficacy over placebo was shown for all examined substances in a recent network-metanalysis involving over 116000 patients [8]. While minor differences between substances exist [7, 8], overall response rates are comparable between 50 to 75% [7, 6]. Patients that don’t show response (commonly defined as reduction of symptom severity > 50% on a standardised scale [7]) should receive a change in treatment strategy.

The timeframe for onset of antidepressant response is often within the first weeks of treatment. A metanalysis including 17 trials with more than 14000 patients showed early improvement (>20 % decrease in symptom severity after 2 weeks) to be a sensitive predictor of later response [9]. The time until full antidepressant effect varies more widely between patients. A secondary analysis of the GENDEP study of 811 patients concluded, that “the eventual outcome of 12-week antidepressant treatment can be accurately predicted only after 8 weeks” [10].

Guidelines try to balance these timeframes in order to avoid unnecessary prolongation of ineffective treatments on one hand and not change the treatment strategy too often and early on the other hand [7]. The current German guidelines recommend evaluation and possible adaptation of treatment strategy after 4 weeks for adults [6, p. 78]. Studies have investigated, whether patients who don’t show early improvement as per the definition above benefit from an earlier change of pharmacological strategy, but overall there is currently insufficient evidence to clearly show a benefit of early medication change for patients without early improvement [7].

This thesis uses secondary analysis of the data from the “EMC trial” [11, 12] in order to attempt a more detailed description and – where possible – prediction of early improvement during an antidepressant treatment course and later progression to response or remission. If there are identifiable and/or predictable clusters of patients that share common response patterns, these clusters might allow a more fine-grained decision-making process compared to the early improvement criterium. These decisions could ultimately lead to clearer recommendations of which patients benefit from early medication change or prolonged treatment continuation.

Background

Diagnosis of Major Depressive Disorder

According to the DSM-V, MDD is a mental disorder characterised by depressed mood being present nearly every day for most of the day and loss of interest or pleasure. Other symptoms include weight loss, insomnia or hypersomnia, psychomotor agitation or retardation, fatigue, feelings of worthlessness or guilt, diminished ability to concentrate or think, indecisiveness and thoughts about death or suicidal ideation [13].

The ICD-10 defines a major depressive episode by patients having more than four of ten symptoms for a duration of at least two weeks. Three of the ten symptoms are main symptoms, of which at least two must be present. The main symptoms are depressed mood, loss of interest or pleasure and loss of energy. Other symptoms are reduced concentration, reduced feeling of self-worth, feelings of guilt, negative outlook for the future, suicidal ideation, disturbed sleep and diminished appetite [14].

Major depressive episodes can occur on their own or repeatedly over the course of a patient’s life. Differential diagnostic includes, among others, bipolar disorder and psychotic disorders [13, 14]. For bipolar disorder to be present, a single manic or hypomanic episode in the history of a patient with a current depressive episode is sufficient [13, 14]. For the purpose of this thesis, MDD, depression and unipolar depression will be used synonymously.

Symptom severity questionnaires

Severity of depression is commonly measured via sum scores of individual symptoms (see [15] for a critique of this practice). Commonly used scales include the clinician rated Hamilton Depression Rating Scale (HDRS or HAMD) [16], the clinician rated Montgomery Åsberg Depression Rating Scale (MADRS) [17], the self-rated Becks Depression Inventory (BDI) [18] and the clinician- or self-rated Inventory of depressive Symptomatology (IDS-30) [19].

Patients are commonly classified as depressed or not depressed based on a cut-off on these sum scales. Different degrees of severity are also commonly distinguished by thresholds on the sum scales [15]. Remission is commonly defined as patients falling below a cut-off on the sum-scales after they have been classified as depressed before [20]. Response is commonly defined as a reduction of sum-scale value by 50% or more (see [12] for an example). This 50% reduction will be synonymously referred to as “traditional response criterium”, “response criterium” or simply “response” for the purpose of this thesis.

Antidepressant treatment strategies

Antidepressants are an important part of treatment strategy for patients with MDD. Current guidelines recommend pharmacological treatment as possible alternative for MDD of medium and strong severity [6]. Antidepressants are being recommended as first-line treatment if pharmacological treatment is chosen [7]. Efficacy of all substance classes of antidepressant medication over placebo has been shown in a recent network-metanalysis involving over 116000 patients [8]. While minor differences between antidepressant substances exist [7, 8], overall response rates are comparable between 50 to 75% [7, 6]. In practice, antidepressants are thus often chosen based on their risk and side-effect profiles [6]. Clinical practice guides commonly recommend starting with selective serotonin reuptake inhibitors (SSRI) like Escitalopram or Sertraline, since they show a good risk-benefit profile [21]. Both clinical practice guides and current guidelines recommend starting antidepressants on a low starting dose and escalating as fast as possible (based on individual patient acceptability and safety) to a standard dose [6, 21]. An appropriate dosage level can further be asserted with therapeutic drug monitoring [6, 21].

The timeframe after reaching a sufficient dosage until treatment evaluation is subject to scientific discussion. Current German guidelines recommend continuation for a duration of 4 weeks (6 weeks for elderly patients). In case of response (defined as a decrease of 50% on a standardised symptom severity scale) the medication should be continued further, until remission (defined as absence of depressive symptoms, e.g. a HAMD score <= 7) is achieved. If there is no response at the evaluation timepoint, guidelines recommend a switch in pharmacological strategy [6]. In practice, a switch to a combination of SSRI and Mirtazapine or a switch to a SSNRI (e.g. Venlafaxine) is common [21]. Alternative strategies include augmentation with lithium or second-generation antipsychotics [6, 7]. It is unclear, whether an earlier evaluation using a different metric is beneficial. Guidelines discuss the early improvement criterium as possible alternative [7, 6, 21].

Early Improvement

As briefly explained in the introduction, patients with MDD often show an early onset of treatment effect to antidepressants. Nierenberg et al. (2000) evaluated 182 outpatients with MDD who responded to fluoxetine treatment [22]. They defined onset of response as a 30% decrease in HAMD score that persisted and led to a decrease of HAMD score over 50% by week 8. With this design, 55.5%, 80.2% and 89.5% (cumulatively) of responders had shown initial response by week 2, 4 or 6 respectively. Szegedi et al. (2003) defined early improvement as 20% decrease of HAMD-Score and showed that the majority of patients treated with mirtazapine (72.7% of 109 patients) or paroxetine (64.9% of 103 patients) showed early improvement within 2 weeks and that this early improvement was a sensitive predictor for later stable response with sensitivity of 0.97/0.91 and specificity of 0.53/0.50 for mirtazapine and paroxetine, respectively [23]. Further investigation into the time course of onset was done by Katz et al. (2004) [24]. 70 patients were randomly assigned to receive 6 weeks of paroxetine, desipramine or placebo. By week 2, there were significant between group differences in symptoms of motor retardation, hostility and depression severity. Most importantly for the topic of this thesis, “the global severity measure […] detected differences between paroxetine responders and nonresponders [sic] as early as 1 week, and this difference was sustained at 2 weeks”.

These results were replicated in a meta-analysis by Szegedi et al. (2009) of over 6000 patients with MDD that received mirtazapine compared to active controls or placebo [25]. Early improvement (> 20% reduction of HAMD-Score) predicted stable response and stable remission with a sensitivity of 81%/87% respectively. Positive predictive values and specificity were comparatively low, so the authors suggested non-improvement after 2 weeks as possible trigger for making early treatment adaptations. A more recent meta-analysis by Wagner et al. (2017) with over 14000 patients assessed the predictive value of the early improvement criterium and found a sensitivity of 85% (95%-CI: 84.3 to 85.7) and specificity of 54% (95%-CI: 53.1 to 54.9) [9].

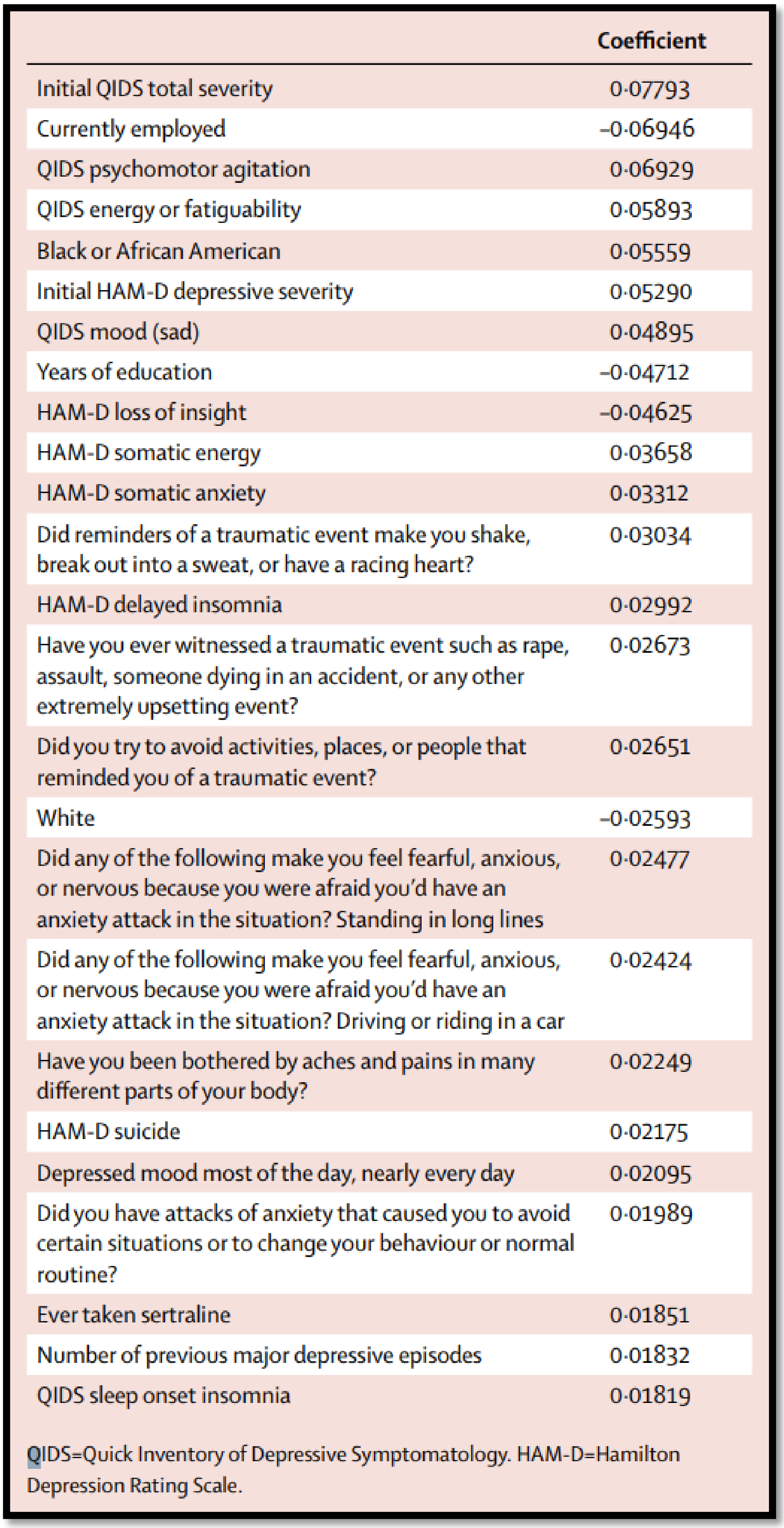
Investigations whether early non-improvement was a suitable trigger for an early medication change (EMC) strategy were done and showed mixed results. Nakajima et al. (2011) treated patients with 50 mg/d sertraline and randomized a total of 41 patients who showed non-improvement after 2 weeks into a group that continued to receive sertraline (n=20) and a group that was switched to paroxetine (n=21) [26]. The switching group had significantly more responders, remitters as well as significantly higher reduction in symptom severity. The larger “EMC trial” [12, 11], that is the main data source for the secondary analysis in this thesis and explained in much more detail later, found no significant differences in outcomes of 192 patients that were randomized into an early medication change and a continuation group.

Predictors of response

In addition to the early improvement criterium, a multitude of demographic and clinical markers have been investigated as to their predictive value for later response or remission.

Many of these analyses were based on the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) trial [27], a large multicentre trial that included over 4000 patients. Rush et al. (2008) looked at a subset of 727 patients that did not show remission or where intolerant to the trials first line treatment with citalopram and analysed for predictors of response to 3 differing second line medications with bupropion, sertraline or venlafaxine [28]. No significant differences for predicting one medications efficacy over the others were found, but several overall predictors of efficacy could be identified: Remission was more likely for employed (vs. unemployed) and married/cohabiting (vs. not cohabiting) patients. Additional predictors were insurance status, previous suicide attempts, DSM-IV axis 1 comorbidity, anxious or melancholic depression characteristics and overall symptom severity. As a more focused analysis for older patients, Kozel et al. (2008) analysed 574 depressed patients over age 55 from the STAR*D trial and compared patients with late onset of MDD (first depressive episode after age 55) with those with earlier onset [29]. No significant differences in remission-rates or time to remission between the onset groups were found. Drago et Serretti (2011) compared the predictor results from the STAR*D trial with an Italian sample of 236 patients [30]. Sociodemographic predictors of remission "included the simultaneous presence of: higher education, higher income, not living alone, and with a good employment status". Nierenberg et al. (2000) investigated predictors for the time to initial response in 182 patients with MDD that responded to fluoxetine [22]. They found, "[n]either demographics (age and sex) nor characteristics of depression (duration of current episode, number of episodes, age at onset of first episode, and baseline score on Hamilton depression scale) predicted time to initial response or time to response by Cox regression analysis for proportional hazards". Comparing responders and non-responders to fluoxetine, some significant differences were found. Non-Responders were more likely to be unemployed and had slightly higher baseline HAMD sum scores.

In addition to just correlating clinical and demographic data, Chekroud et al. (2016) [31] built a full machine learning model to predict remission from treatment with escitalopram using the STAR*D dataset and validated their model on an external dataset from the COMED study (see [32] for details). They managed to achieve an accuracy of 64.6% in the internal cross-validation, significantly over chance. In order to build the model, the authors used all variables that were overlapping in the STAR*D and the validation dataset and selected the top 25 predictors by elastic net regularisation. An overview of the predictors is found in Figure 1. The top predictors included symptom severity, specific clinical symptoms and demographic markers. Accuracy in external validation was 59.6% ($p = 0.043$) for patients treated with escitalopram, 59.7% ($p=0.023$) for patients treated with escitalopram and bupropion and 51.4% ($p=0.53$) for patients treated with venlafaxine and mirtazapine, which suggests predictive value being (partially) specific to the mechanism of antidepressant action.



	Coefficient
Initial QIDS total severity	0·07793
Currently employed	–0·06946
QIDS psychomotor agitation	0·06929
QIDS energy or fatiguability	0·05893
Black or African American	0·05559
Initial HAM-D depressive severity	0·05290
QIDS mood (sad)	0·04895
Years of education	–0·04712
HAM-D loss of insight	–0·04625
HAM-D somatic energy	0·03658
HAM-D somatic anxiety	0·03312
Did reminders of a traumatic event make you shake, break out into a sweat, or have a racing heart?	0·03034
HAM-D delayed insomnia	0·02992
Have you ever witnessed a traumatic event such as rape, assault, someone dying in an accident, or any other extremely upsetting event?	0·02673
Did you try to avoid activities, places, or people that reminded you of a traumatic event?	0·02651
White	–0·02593
Did any of the following make you feel fearful, anxious, or nervous because you were afraid you'd have an anxiety attack in the situation? Standing in long lines	0·02477
Did any of the following make you feel fearful, anxious, or nervous because you were afraid you'd have an anxiety attack in the situation? Driving or riding in a car	0·02424
Have you been bothered by aches and pains in many different parts of your body?	0·02249
HAM-D suicide	0·02175
Depressed mood most of the day, nearly every day	0·02095
Did you have attacks of anxiety that caused you to avoid certain situations or to change your behaviour or normal routine?	0·01989
Ever taken sertraline	0·01851
Number of previous major depressive episodes	0·01832
QIDS sleep onset insomnia	0·01819

QIDS=Quick Inventory of Depressive Symptomatology. HAM-D=Hamilton Depression Rating Scale.

*Figure 1: Top 25 predictors of remission from depression after escitalopram from elastic net regularization in the STAR*D dataset. Graph taken from [31].*

Paul et al. (2019) [33] tried to predict classes of treatment response developed on the MARS (see [34]) and GENDEP (see [35]) samples with a random forest algorithm. For details of the target treatment response classes, see page 14. As prediction variables, they used sociodemographic, psychiatric and family history, vital signs and baseline laboratory data, life events, baseline psychopathology and personality items. As model extensions, they added baseline HAMD single items and HAMD early partial response after 2 weeks. The predictive models achieved "classification accuracies between 75 and 95.2%". It should be noted though, that these accuracies reflect a prediction on whether a given patient belongs to a single cluster or not and not a prediction of which cluster a given patient belongs to, so much higher accuracies should be expected due to the higher zero information rate (meaning the highest accuracy achievable by predicting only one class).

Patterns of treatment response over time

In a secondary analysis of the GENDEP study (see [35] for details), Uher et al. (2001) applied longitudinal latent class analysis of relative symptom severity scores to 811 depressed patients that received escitalopram or nortriptyline for 12 weeks [10]. The number of groups was evaluated using the Bayesian information criterion, based on which the authors selected a model of 9 trajectories (see Figure 2). A key result of this analysis was that, while early improvements in depression severity are commonly maintained and lead to response, there are classes that show response with a significant delay. According to the authors, the "eventual outcome of 12-week-antidepressant treatment can be accurately predicted only after 8 weeks".

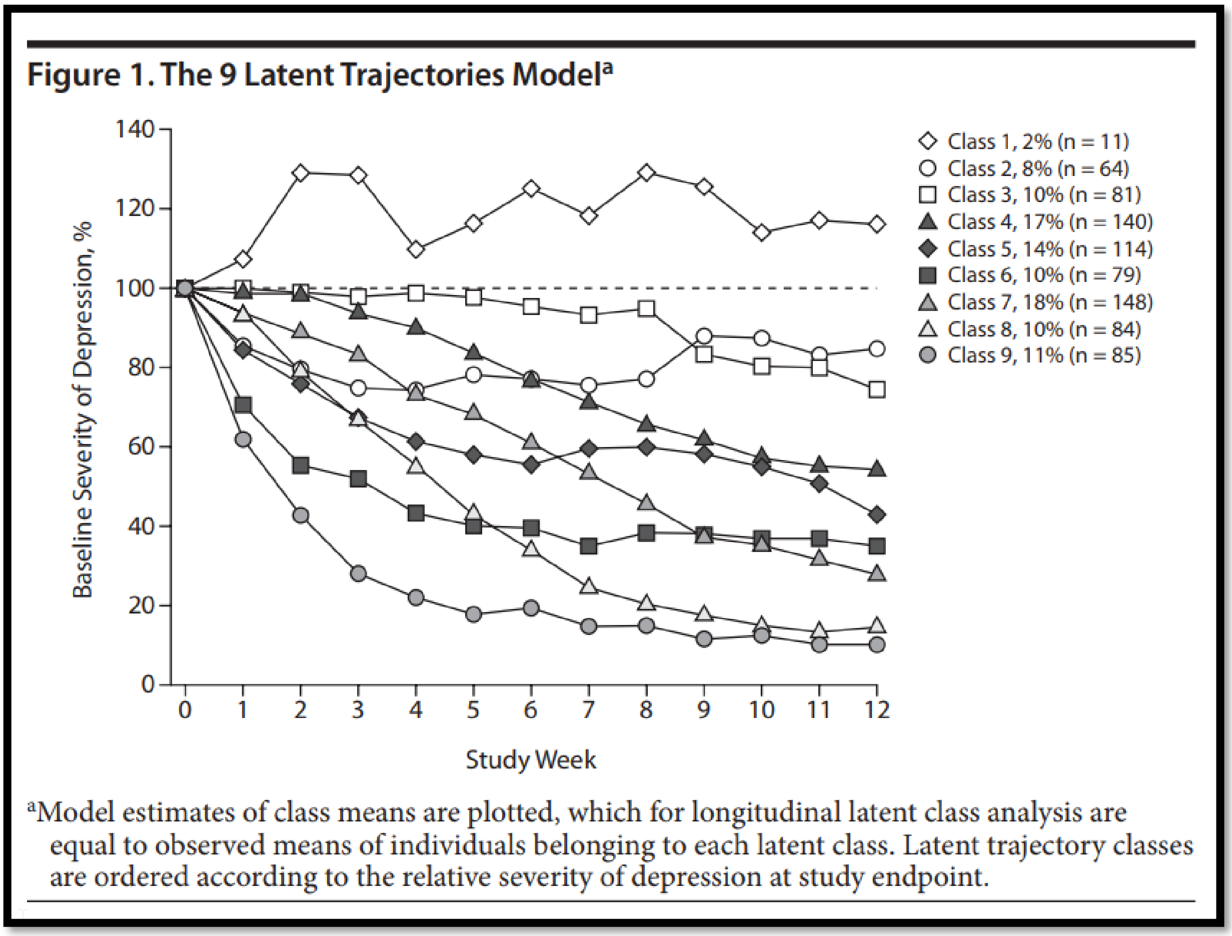


Figure 1. The 9 Latent Trajectories Model[a]

[a]Model estimates of class means are plotted, which for longitudinal latent class analysis are equal to observed means of individuals belonging to each latent class. Latent trajectory classes are ordered according to the relative severity of depression at study endpoint.

Figure 2: A 9 class longitudinal latent trajectories model of depressed patients from a secondary analysis of the GENDEP study. Graph taken from [10].

Another attempt of identifying clusters of treatment response, Paul et al. (2019) [33] created clusters from 809 patients of the MARS study (see [34] for details) and validated these clusters based on a holdout sample from the MARS study (n=236) and patients from the GENDEP (see [35]) study (n=826). Clusters were created on logarithmically transformed weekly HAMD sum scores over up to 16 weeks with a mixed model approach with the number of clusters being assessed by the integrated completed likelihood criterium. With this approach, 7 treatment response classes were identified. These are shown in Figure 3.

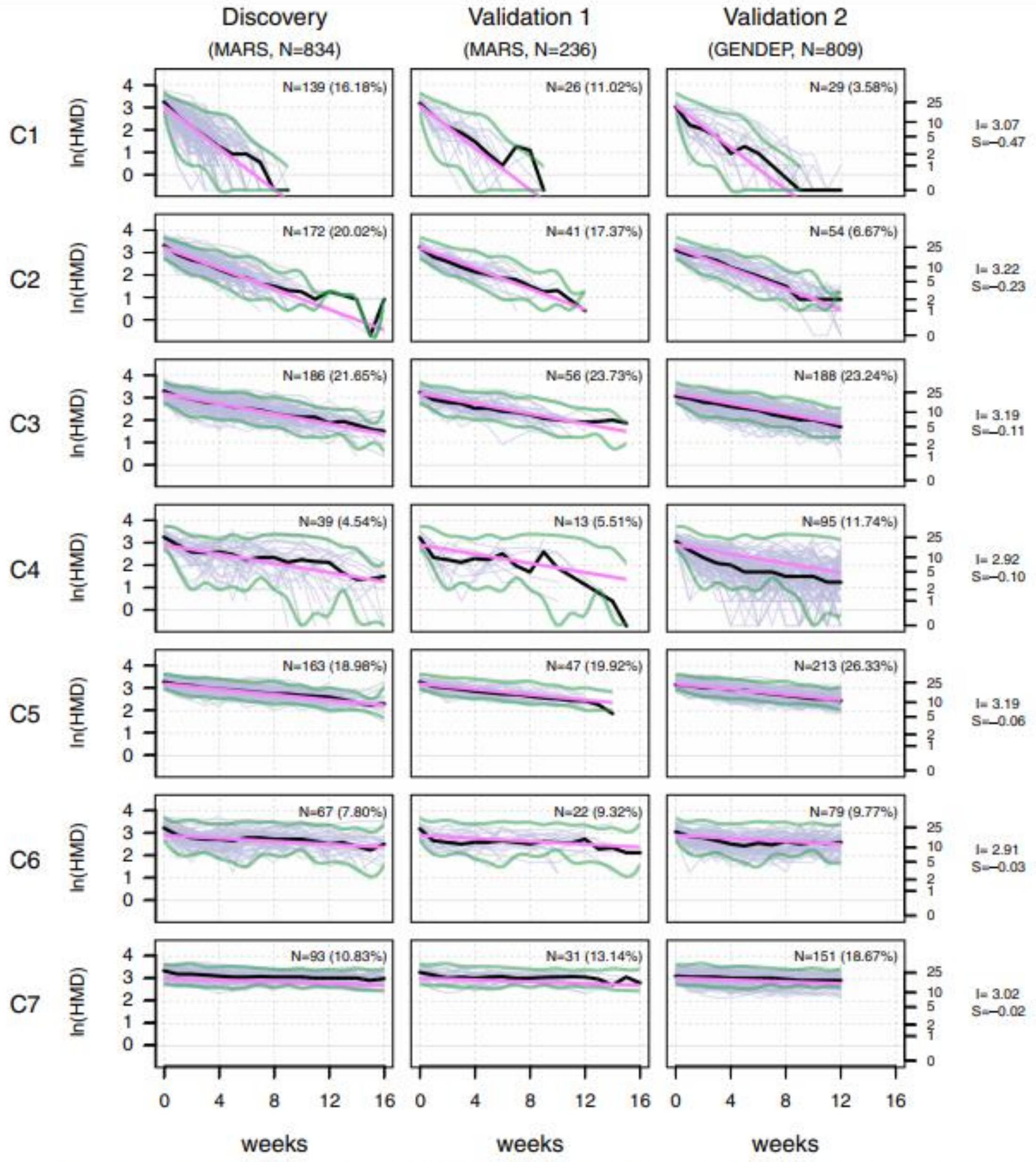


Fig. 1 Resulting cluster shape characteristics and underlying natural logarithm-transformed HAM-D courses for the discovery sample and both validation samples. *X-axis:* observation time in weeks; *Y-axis:* natural logarithm-transformed HAM-D values (purple: raw values, black: cluster-specific median, pink: model-based linear fit). Slope and intercept values of all clusters are given on the right. Clusters are sorted from C1 to C7 according to the cluster-specific slope. Absolute and relative cluster sizes in all samples are given within the subplots. Green borders represent the limits in which 95% of HAM-D values of the discovery sample were contained. These were transferred to columns 2 and 3 to allow for comparison with the validation samples. S slope, I intercept, ln natural logarithm-transformed

Figure 3: A 7 class mixed model on logarithmic HAMD sum scores based on the MARS and validated by a holdout set of MARS as well as the GENDEP study datasets. Graph and description taken from [33].

Rationale

Due to the possible inter-individual variety in antidepressant treatment response, finding an optimal timepoint for treatment evaluation is difficult. The approach of evaluation after 4 weeks that is currently recommended by German guidelines [6] is insufficient to be used for all patients, since some patients will show response much later in the treatment course [10]. At the same time, prolonging the treatment over a longer amount of time would simultaneously prolong the duration of the disease for a large part of patient population (25-50% [7, 6]), that doesn't show response to (at least) the first antidepressant substance.

Balancing the opposing needs for both these patient groups in an optimal way has an additional difficulty in the operationalisation of response. By utilizing standardised sum scores and their relative change as the only criterium, information about patients with depression is reduced to the level of current symptom severity, while other clinical characteristics are disregarded. This approach seems insufficient since previous data shows the predictive value of both demographic and clinical history factors (See page 11).

Optimization for the overarching problem of individual treatment response evaluation is attempted in several ways. One of those is to define separate evaluation timepoints for specific patient groups. One example would be the evaluation after 6 weeks for elderly patients that is recommended in current German guidelines [6]. Another, more recent, attempt is to evaluate patients after 2 weeks based on the early improvement criterium (see page 10). Since the early improvement criterium is quite sensitive, patients without said early improvement have a low chance of showing response later in the course of treatment. Thus, early medication change (EMC) might be beneficial for this group of patients. As explained above, it is currently still unclear, whether this approach is clinically preferable.

This thesis attempts to generate new insights for future approaches to the response evaluation problem explained above by utilizing machine learning techniques (explained in detail below) to investigate the first 4 weeks of treatment – the timeframe until traditional response evaluation would be performed. The goal is to generate data driven hypotheses that can be further evaluated for use in clinical decision making while at the same time providing and evaluating the necessary algorithmic tools to facilitate their future investigation.

For these hypotheses to provide a possible benefit over current clinical decision making, they must either incorporate an intervention (e.g. early medication change) before the traditional evaluation at week 4 or describe a subset of the population for which treatment evaluation at week 4 would be too early. In order to find data-driven hypotheses to answer these problems, two sources of information are combined. First, instead of looking at cut-offs at a certain timepoint, the early response to treatment is investigated over time. This could enable a more fine-grained classification of early treatment response and thus facilitate identification of population subsets for which separate evaluation timepoints might be appropriate. Second, the predictive aspects from patient history and symptom severity scores are combined with the time-course analysis in order to predict later treatment outcome before the fact, where possible. This could then enable early intervention for a subset of patients for whom accurate prediction is possible and intervention is beneficial.

Common Methods

Data Source – The EMC Trial

The main data source for the analysis in this thesis is the “Randomised clinical trial comparing early medication change (EMC) strategy with treatment as usual (TAU) in patients with Major Depressive Disorder” or “EMC trial”. For further information on topics in this chapter, please refer to published information on the study protocol [12]. The EMC trial’s primary objective was to compare the effectiveness of an early medication change regimen as compared to treatment as usual. Early medication change (EMC) in the context of the EMC trial refers to switching the antidepressant substance after 14 days in patients that did not show early improvement as opposed to treatment as usual (TAU), which involves continuation of the same antidepressant for a total of 28 days.

The EMC trial used a three-level randomization process that is summarized in Figure 4.

On level 1, all patients received Escitalopram (ESC) for 14 days. Patients showing early improvement (decrease < 20% in HAMD17) between day 0 and day 14 continued to receive Escitalopram for another 14 days and were taken to level 2. Patients not showing improvement were randomized into an EMC group (EMC1) receiving Venlafaxine (VEN) and a TAU group (TAU1) continuing Escitalopram for 14 days each. The TAU group was rated as responders or non-responders based on the response criterium (decrease < 50% in HAMD17) between day 0 and day 28. Responders continued to receive Escitalopram, non-responders received Venlafaxine. The EMC group was again split by the early improvement criterium between day 14 and day 28. Improvers continued to receive Venlafaxine, non-improvers received venlafaxine and lithium.

Patients that showed early improvement in level 1 and received Escitalopram until day 28 were taken into level 2. Patients of this subgroup that showed no response (< 50% HAMD17 decrease) were switched to Venlafaxine. After 14 days of Venlafaxine, the early improvement criterium was evaluated for this timeframe (>20 % HAMD17 decrease between day 28 and day 42). Patients that did show early improvement in this timeframe continued to receive venlafaxine for 14 more days. Patients without early improvement were randomized into the EMC2 and TAU2 groups with EMC2 receiving venlafaxine with lithium augmentation and TAU2 receiving continued venlafaxine for 14 days (until day 56). Patients that did show response on day 28 continued treatment with escitalopram and were taken to level 3.

In level 3, patients had shown a response to Escitalopram on day 28. If there was no further improvement (<20% HAMD17 decrease) between day 28 and day 42, patients were randomised to EMC3 or TAU3 that both lasted 14 days. EMC3 was switched to venlafaxine, TAU3 continued to receive Escitalopram. Patients that did show improvement between day 28 and day 42 continued to receive Escitalopram.

For all levels, patients that showed remission (HAMD 17 absolute score <= 7) were counted as improvers or responders, even if they did not meet the relative criterium. Medication with Escitalopram and Venlafaxine was escalated to the highest tolerable dose or the maximum dose (20 mg/d and 375 mg/d respectively). Lithium dose was adjusted for a plasma level range of 0.6 to 0.8 mmol/l.

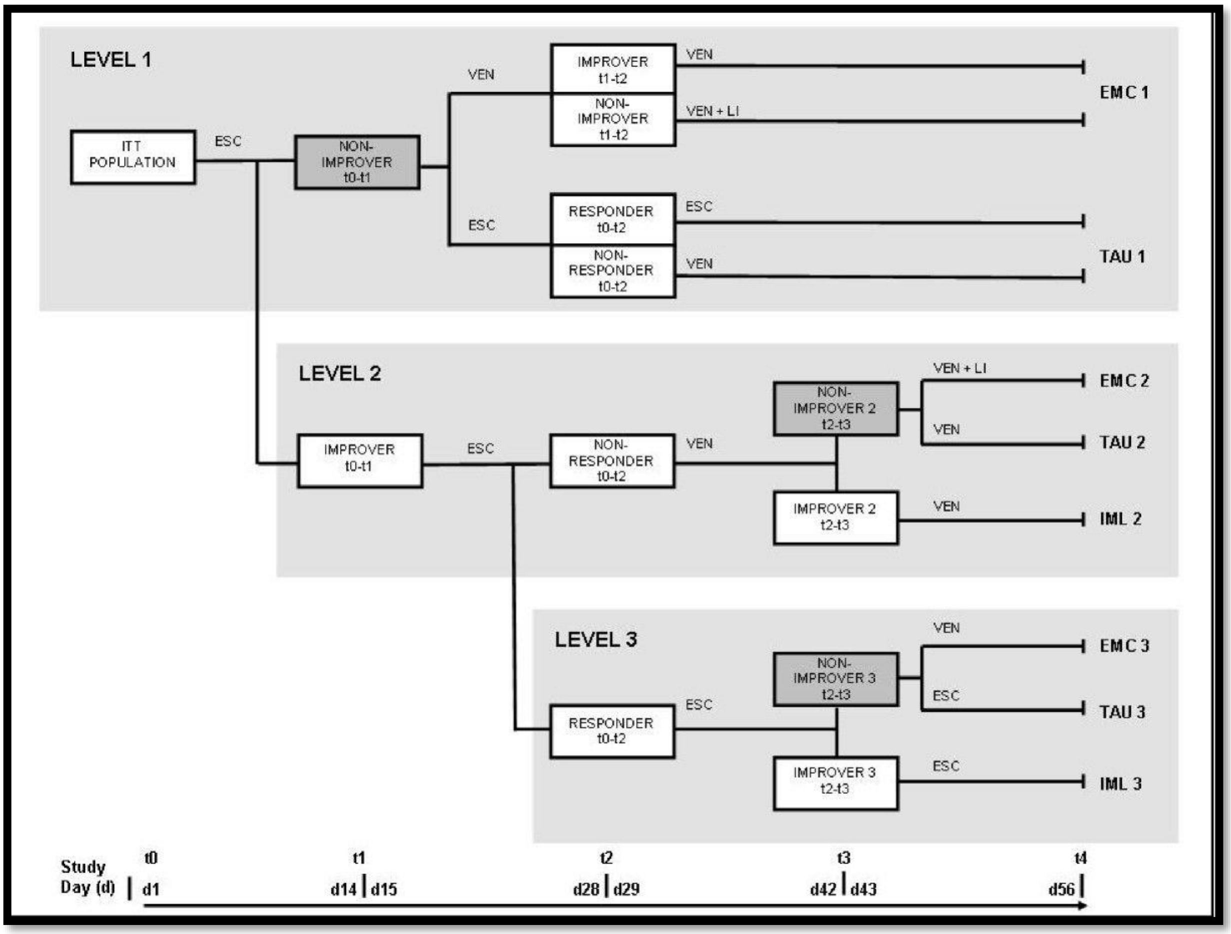


Figure 4: Overview of treatment arms in the EMC trial. Graph taken from [12].

The primary endpoint of the trial was remission with a HAMD17 sum score <= 7 on day 56 which was compared between EMC1 and TAU1 groups. Secondary endpoints were response (HAMD17 decrease >= 50% compared to baseline), absolute change of HAMD17 sum score, response and remission (sum score <= 11) in clinician and self-rated IDS-30 questionnaires.

Inclusion criteria for the EMC trial were patients with moderate MDD (HAMD17 sum score >= 18 points) with an age between 18 and 65 with the first depressive episode occurring before age 60 that understood and signed the informed consent form. For all exclusion criteria, refer to the study protocol [12], some key ones (in the author's opinion) are listed here for summary:

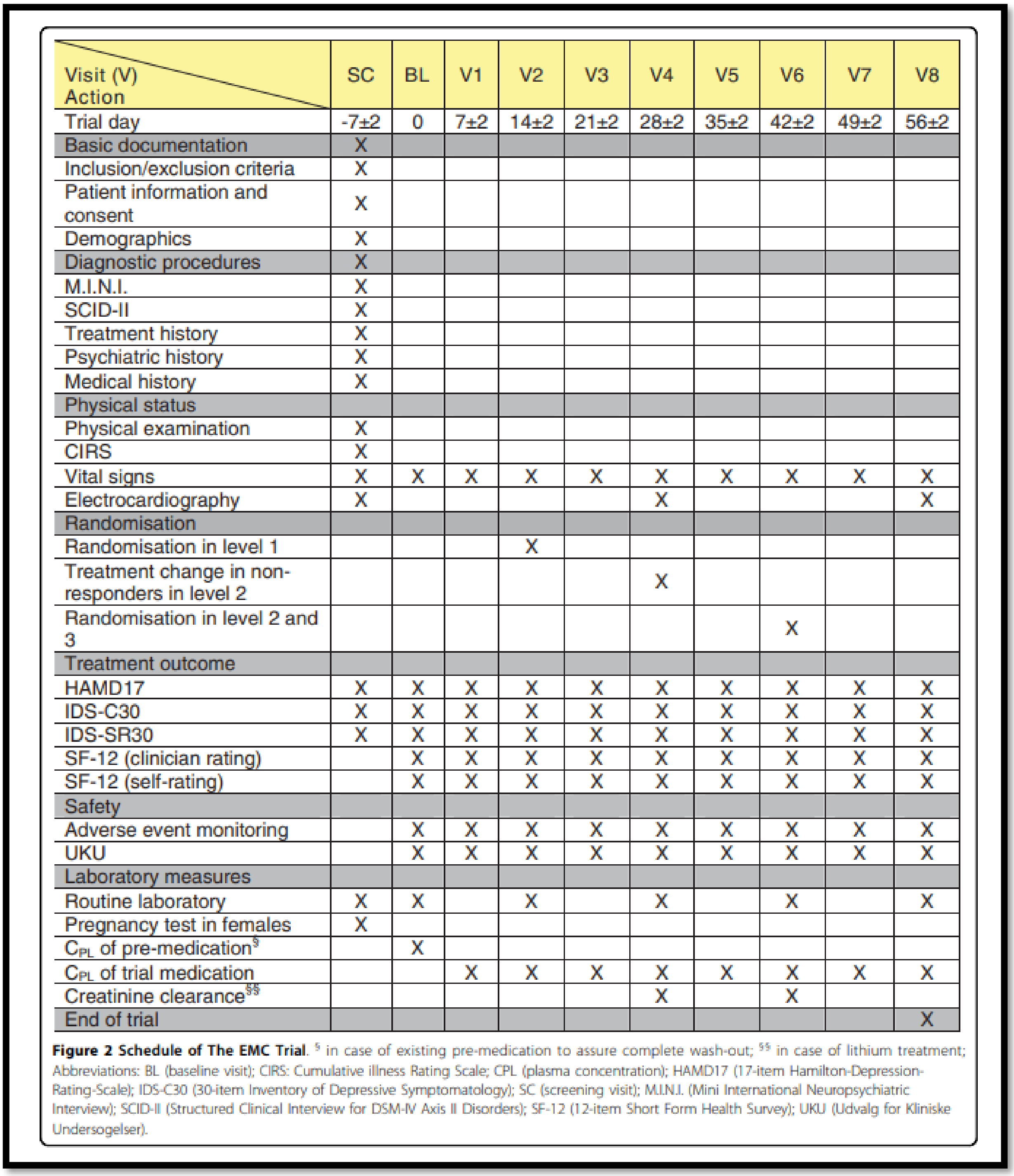
- Necessary intervention outside of protocol treatment because of suicide risk
- Lifetime diagnosis of dementia, schizophrenia, schizoaffective disorder or bipolar disorder
- Current diagnosis of PTSD, OCD, anxiety disorder or eating disorder requiring a non-protocol treatment

- Current substance dependency requiring detoxification
- Depression secondary to organic disorder (e.g. multiple sclerosis)
- Clear history of non-response in the current depressive episode to any protocol medication

Dataset

Patients in the EMC trial were assessed using a variety of psychometric and clinical measurements. For an overview of these measurements, see Figure 5. To summarize for the purpose of this thesis, at the screening timepoint (day -7 +/- 2), demographics, medical and psychiatric history were taken. At the same timepoint, the Mini-International Neuropsychiatric Interview (M.I.N.I.) [36] and the Structured Clinical Interview for DSM-IV Axis II (SCID-II) [37] were conducted. Symptom severity was assessed weekly with the HAMD score and both the clinician and self-rated versions of the IDS.

The dataset available for secondary analysis doesn’t include the complete measurements taken in the EMC trial. The most notable difference for the purpose of this thesis is the availability of itemized HAMD and IDS measurements only until week 2. For week 3 and after, only sum scores are available in the dataset. Since this thesis focuses on the early timeframe of treatment up to week 4 (until traditional evaluation would occur), only patients that have HAMD scores for at least 4 weeks are included for analysis. Population characteristics for this subpopulation are described below.



Visit (V) Action	SC	BL	V1	V2	V3	V4	V5	V6	V7	V8
Trial day	-7±2	0	7±2	14±2	21±2	28±2	35±2	42±2	49±2	56±2
Basic documentation	X									
Inclusion/exclusion criteria	X									
Patient information and consent	X									
Demographics	X									
Diagnostic procedures	X									
M.I.N.I.	X									
SCID-II	X									
Treatment history	X									
Psychiatric history	X									
Medical history	X									
Physical status										
Physical examination	X									
CIRS	X									
Vital signs	X	X	X	X	X	X	X	X	X	X
Electrocardiography	X					X				X
Randomisation										
Randomisation in level 1				X						
Treatment change in non-responders in level 2						X				
Randomisation in level 2 and 3								X		
Treatment outcome										
HAMD17	X	X	X	X	X	X	X	X	X	X
IDS-C30	X	X	X	X	X	X	X	X	X	X
IDS-SR30	X	X	X	X	X	X	X	X	X	X
SF-12 (clinician rating)		X	X	X	X	X	X	X	X	X
SF-12 (self-rating)		X	X	X	X	X	X	X	X	X
Safety										
Adverse event monitoring		X	X	X	X	X	X	X	X	X
UKU		X	X	X	X	X	X	X	X	X
Laboratory measures										
Routine laboratory	X	X		X		X		X		X
Pregnancy test in females	X									
C_{PL} of pre-medication§		X								
C_{PL} of trial medication			X	X	X	X	X	X	X	X
Creatinine clearance§§						X		X		
End of trial										X

Figure 2 Schedule of The EMC Trial. § in case of existing pre-medication to assure complete wash-out; §§ in case of lithium treatment; Abbreviations: BL (baseline visit); CIRS: Cumulative illness Rating Scale; CPL (plasma concentration); HAMD17 (17-item Hamilton-Depression-Rating-Scale); IDS-C30 (30-item Inventory of Depressive Symptomatology); SC (screening visit); M.I.N.I. (Mini International Neuropsychiatric Interview); SCID-II (Structured Clinical Interview for DSM-IV Axis II Disorders); SF-12 (12-item Short Form Health Survey); UKU (Udvalg for Kliniske Undersogelser).

Figure 5: Overview of measurements from the EMC trial. Table and description taken from [12].

Population characteristics

For a description of the EMC trial dataset population, Table 1 and Table 2 give information on sociodemographic and clinical data. For these reports, only patients without missing HAMD-Scores until at least week 4 have been included, since these constitute the dataset that is used for all further analysis in this thesis.

Variable	**Categorical Values**	**n (% of total)**
Total n		766 (100%)
Gender	Male	327 (42.7%)
	Female	438 (57.3%)
Ethnic Group	European	739 (96.6%)
	Asian	7 (0.9%)
	African	6 (0.8%)
	Other	13 (1.7%)
Highest School Degree	None	9 (1.2%)
	Lower secondary school	235 (30.7%)
	Intermediate secondary school	237 (31.0%)
	Advanced technical certificate	93 (12.2%)
	Upper secondary school	185 (24.2%)
	Other	6 (0.8%)
Highest Vocational Degree	None	102 (13.3%)
	Apprenticeship	446 (58.3%)
	Master	20 (2.6%)
	University, College of higher Education	163 (21.3%)
	Vocational College	28 (3.7%)
	Other	6 (0.8%)
Recurrent MDD	First Episode	259 (33.9%)
	Previous Episodes	506 (66.1%)
Age of MDD Onset	Early Onset (before Age 21)	191 (25.0%)
	Middle Onset (Ages 21 to 44)	416 (54.4%)
	Late Onset (After Age 45)	157 (20.5%)

Table 1: Categorical demographic and clinical descriptors of the EMC dataset. Differences in absolute sums are due to missing values, differences in percentage sums are due to rounding error. Only patients without missing HAMD Scores until at least week 4 have been included.

Variable	Mean (95% CI)
Age	40.6 (95% CI 39.8 to 41.4)
Years of Education	13.9 (95% CI 13.7 to 14.2)
Age at MDD onset	32.3 (95% CI 31.4 to 33.2)
Number of previous MDD episodes	2.63 (95% CI 2.31 to 2.96)
Length of current episode (days)	31.8 (95% CI 28.0 to 35.7)
HAMD score at Baseline	22.9 (95% CI 22.6 to 23.2)

Table 2: Continuous demographic and clinical descriptors of the EMC dataset. 95% CI: 95 % confidence interval for the mean. Only patients without missing HAMD Scores until at least week 4 have been included.

Software, source code and open source policy

Code for all analyses in this thesis was written with Python and Open Source Software Libraries. A list of used packages with the corresponding software versions is given in Table 3. The full source-code for all experiments in this thesis is available on request through the author. This code or any part of it is free to use, on the condition, that this thesis is cited as source. The source code has been written with great care, though the author accepts no responsibility or liability for any damages caused by its use. The datasets used for the analysis are proprietary and as such, cannot be shared without prior approval. For scientific inquiries regarding the datasets, please contact the author.

Python Software Package	Version Number
Python [38]	3.6.6
pandas [39]	0.24.1
NumPy [40]	1.16.4
Matplotlib [41]	2.2.2
scikit-learn [42]	0.19.2
SciPy [43]	1.3.1

Table 3: Major software packages used in this thesis. Dependencies of the given software are not mentioned.

Structure of the experiments in this thesis

This thesis has a total of three Experiments. Experiments 1 and 2 stand mostly on their own, while Experiment 3 combines key results from the previous two. While great care has been taken to make this structure easily accessible on a first “top to bottom” readthrough, the author suggests first consulting this thesis’ abstract (see page 96) as well as the experimental summaries (see pages 25, 57 and 76 for experiments 1 through 3 respectively) before continuing. This should help with a clearer understanding of all experimental steps and findings in context.

Results from all three experiments are discussed separately after each experiment. These results are then added to a combined discussion (see page 86). This combined discussion should not be read separately, since aspects from all three individual experimental discussions (see pages 46, 71 and 82 for experiments 1 through 3 respectively) have been presumed as known for its purpose.

Experiment 1 – Clustering early treatment response

Summary of Experiment 1

In Experiment 1, the k-means-algorithm was used in order to identify possible clusters of treatment response until week 4. The time course of response was operationalized by HAMD-Scores relative to baseline. Different numbers of clusters (k) were calculated and goodness of fit data and clinical interpretation were discussed in order to identify candidates for further investigation. For k=5, the goodness of fit data showed decent fit and the cluster structure (See Figure 22) was suggestive of a clinical interpretation based around the traditional early improvement (20% decrease) and response criteria (50% decrease).

The cluster structure suggests that patients initially fall into the categories of “Early Improvement”, “Early Non-Improvement” and “Early Response”. Patients that show “Early Non-Improvement” can be consecutively differentiated in patients showing “Delayed Improvement” or “Non-Improvement”. Patients showing “Early Improvement” can be consecutively separated into patients with “Early Improvement with Response” and “Early Improvement without Response”.

Aim

The time course of early treatment response is a possible source of benefits for clinical decision making (See page 16). Limited prediction being possible with the early improvement criterium shows predictive (regarding the outcome of the entire course of treatment) information isn’t just contained in the level of symptom severity at week 4 but also the levels at previous timesteps. The underlying aim for this experiment is to find latent information useful for prediction by classifying the early timeframe of treatment in more detail. In order to facilitate this, patients with similar patterns of symptom severity over time will be grouped together into sets of distinct clusters. These cluster structures will then be evaluated twofold. Firstly, it will be investigated whether they are a mathematically valid (good measures of fitness) and generalizable (to an external dataset) description of the possible patterns of early treatment response. In case this validity is shown, the usefulness of the cluster structure is further investigated in a second step. For this, clinical interpretability is one of the key characteristics that should be discussed in detail, since possible interventions will have to be derived based on clinical interpretations of the cluster structure.

Methods

Design Summary

The relative HAMD sum scores for the first 4 weeks for patients from the EMC dataset are being grouped into clusters by using the k-means clustering algorithm. The fit for a given number of clusters k is assessed by the elbow-method and the silhouette score. The algorithm and fitness measures are explained below. The trained k-means-clustering algorithm is then used to classify an external dataset, which is described in more detail below, for validation. The fitness measures are repeated on the validation set to check for generalization of the algorithm.

K-Means-Algorithm

The k-means-algorithm is a cluster analysis technique that fits a set of observations to a given number of clusters (k). Since the target clusters aren't known to begin with, this is an unsupervised machine learning technique. As starting condition, the algorithm takes k-observations from the set of all observations. This choice is either made at random or following certain rules (e.g. maximal distance between observations). All remaining observations are then assigned to the closest of the initial k observations as measured by an arbitrary distance-measure, most commonly Euclidian distance. This assignment defines k clusters with each observation belonging to one cluster. In the iteration part of the algorithm, the mean value of all observations belonging to a single cluster is calculated. This step is repeated for all k clusters, resulting in k cluster mean values. After this step, all observations are assigned to the cluster mean closest to them, again resulting in k clusters. The iteration part is repeated, until the cluster assignments no longer change during iterations. Using the trained algorithm to classify new data can be done by assigning a new datapoint to the closest cluster mean without updating the mean values. [44] This gives the algorithm several properties worthy of pointing out. Since all observations are assigned to a cluster, there are no "outlier" observations, that will not be assigned to a cluster. The algorithm is also dependent on the starting condition: a different set of initial observations will possibly result in a different result of clusters. This is especially relevant, if the starting condition isn't dependent on a set of rules, but a random process instead.

For this specific experiment, the algorithm was set up to start with a random starting condition. The random number generator was set up with a "seed" to make results repeatable. A "seed" is a condition for the random number generator that ensures all (pseudo-)random numbers being the same on every run of the code. The distance measure chosen was the Euclidian distance, which is considered the default. Since no properties of the data analysed here indicate the need for a non-default distance measure and conformity to the default leads to better comparability, this choice is – in the authors opinion – justified.

Number of clusters and measurements of fitness

The number of clusters (k) is an algorithmic parameter that is predetermined in the k-means-algorithm, but there currently is no theoretical justification for imposing a given number of clusters. Therefore, the clustering process was repeated for any number of clusters between 2 and 10. The maximum of 10 clusters was arbitrarily imposed so clusters could remain useful to clinical interpretation. Interpretability suffers with increased number of clusters, mostly because the average number of patients per cluster decreases. At the same time, more clusters lead to increasingly smaller differences between the clusters, that are then obviously less practically and/or scientifically relevant. By repeating the calculations for different cluster numbers, the need arises to find one or more numbers that have the best "fit" to the data. In this experiment, well established methods for determining fitness were used, namely the so-called elbow method and the silhouette score.

The elbow method relies on the fact, that a higher number of clusters (k) always leads to a higher degree of inter-observation-variance explained by the clusters. As obvious extremes, only 1 cluster explains 0% of variance and a single cluster for each observation explains 100% of variance. The graph of the explained variance dependent on the number of clusters can then be used to find the fitting number of clusters by looking for a flattening of the graph (the so called "elbow") [45].

The silhouette score is commonly used to determine an appropriate k. This method scores every observation for its fit to the cluster it was assigned to. This is done by calculating the average distance between the observation and all other observations of the cluster it was assigned to and subtracting the average distance from all observations of the closest cluster it was not assigned to (the “second best” choice for the cluster) and normalizing the score to a measure from -1 (maximum outlier) to 1 (perfect fit) [46]. Graphing these individual scores in a sorted and cluster-grouped line graph is the so called “silhouette plot”. This plot can be used to visually identify cluster fitness by visible cues such as similar cluster sizes, small outlier (negative) silhouettes as well as smooth and/or convex silhouette shape. Since these cues can be used to assess cluster fitness independent of finding an appropriate k, these are also appropriate visual interpretation aids for use in validation datasets. Calculating the average of all observation scores and graphing it dependent on the number of clusters gives the information on an appropriate cluster number. Both the global maximum as well as outliers from an interpolated graph should be investigated closer. In order to find the latter, for each k between 3 and 9 a linear interpolation of the silhouette scores from k-1 and k+1 is calculated. The differences from the observed values are plotted as residuals and local maxima on this residual graph are candidates for further inspection.

Both the elbow method and the silhouette method are subject to researcher’s interpretation, especially in cases where differences between values of k aren’t pronounced. Dependent on both the goodness of fit measures and theoretical context, multiple k might be chosen as appropriate. Detailed reasoning for candidate k values is therefore given in the discussion section for this experiment.

Relative vs. Absolute HAMD score and outlier handling

By choosing the relative HAMD score as opposed to the absolute values, the dimensionality of the clustering problem can be reduced from 5 to 4 (since all baseline observations are equal to the value 1.0 now) and clusters are easier to interpret clinically. This is especially true regarding both the early improvement criterium and the response measure, that are defined relatively as a decrease of 20% or 50% compared to baseline respectively.

The use of the relative score creates an opportunity for outliers, since there is no value range like there is for the absolute score. These outliers might be assigned to clusters with only a single observation, which can lead to significant problems in choosing an appropriate number of clusters. For that reason, observations that are assigned single item clusters at any experimental stage are manually identified and excluded before all calculations are repeated. Other handling of outliers is not applied.

Missing value handling

The k-means-algorithm as described above can’t natively handle missing values. Thus, a decision on how to handle missing values needs to be made. Any interpolation of missing values (e.g. linear interpolation) might directly influence the aim of this experiment, the cluster-structure. For this reason, all patients with missing values in the variables used for clustering were excluded from analysis.

Cluster numeration

In order to make cluster numbers between different k easier to interpret and reference, cluster numbers are reassigned after the calculations. The reassignment is based on the mean relative HAMD score after 4 weeks with cluster 0 having the lowest and cluster k the highest mean score after reassignment. This step doesn’t change any properties of the clusters or the resulting calculations and is only described for completeness.

Validation data

The k-means classifiers for different k are trained on the EMC dataset as described above. As additional test to the generalization of the classifiers, previously unseen patients from a validation dataset are assigned to clusters and silhouette scores within the validation dataset are calculated as described above. In case of poor generalization an increase in observations with negative silhouette scores and a decrease of average silhouette scores can be expected. Additionally, relative cluster sizes changing widely between training and validation are a sign of poor generalization.

For validation, the “Study 831” dataset from the Psychiatric Hospital, University of Zürich was kindly provided by Prof. Dr. Stassen. This dataset has a total of 1645 patients with HAMD scores over a period of 35 days. HAMD measurements were taken on days 0, 3, 7, 10, 14, 21, 28 and 35. In many cases, there were differences between theoretical and real measurement timing. These differences were provided with the dataset. Further details about the study protocol(s) were not provided because they are inconsequential for the purpose of validation in this context.

In order to use the Study 831 dataset for validation purposes, measurements from days 0, 7, 14, 21 and 28 were selected as variables. We allowed for differences of measurement timing of +/- 2 days resulting in timings being the exact same as for the EMC dataset. All patients with timings differing further and patients with missing values in the selected variables were excluded leaving a total n of 1028 for validation.

Descriptive Reporting

Since the clusters are generated from only the relative HAMD scores, there can be between-cluster differences in sociodemographic and clinical variables. Therefore, reports on the descriptive variables (see Table 1 and Table 2) are generated for all groups of clusters. For categorical variables, the statistical significance of the group differences under the null hypothesis of uniform distribution is estimated with the Chi-Squared-Test. For continuous variables, significance is estimated with a one-way ANOVA.

Results

Number of observations

The total n after eliminating patients with missing values is 766. One additional patient was excluded as outlier because he was classified into a single-element-cluster, so the total n for final calculations was 765. The mean values for the relative HAMD-Scores were approximately 0.68 after 1 week, 0.60 after 2 weeks, 0.57 after 3 weeks and 0.53 after 4 weeks.

Elbow method for estimation of cluster number

The percentage of variance that is explained for the clusters resulting from k-means-clustering for a given number of clusters (k) was calculated for k between 2 and 10 as explained in the method section. The resulting graph is shown Figure 6. The graph doesn’t show a single obvious flattening, so the elbow method is ambiguous. Possible candidates for k given by this method are 3, 4, and 5. This is concluded from the graph showing visible flattening after each of those numbers. After 5 clusters, the explained variance grows almost linear, so no candidates are found in this range.

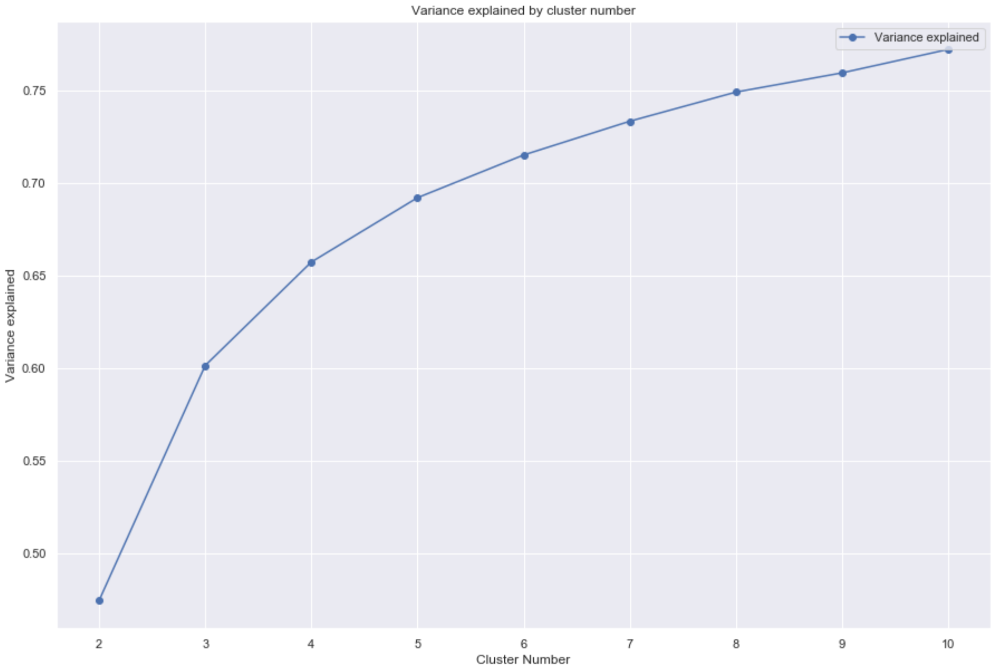


Figure 6: Variance explained by the clusters resulting from k-means-clustering for a given number of clusters k

Average silhouette score for estimation of cluster number

The average silhouette scores for the clusters resulting from k-means-clustering for a given number of clusters (k) was calculated for k between 2 and 10 as explained in the method section. The resulting graph is shown in Figure 7. The graph shows a global maximum at 2 clusters with a mostly monotonous decrease shaped roughly like an exponential decay. There is a positive outlier from the exponential decay shape of the graph at 5 clusters. The silhouette score at this outlier doesn’t quite reach the amount of the score at 4 clusters, so it can’t be considered a local maximum.

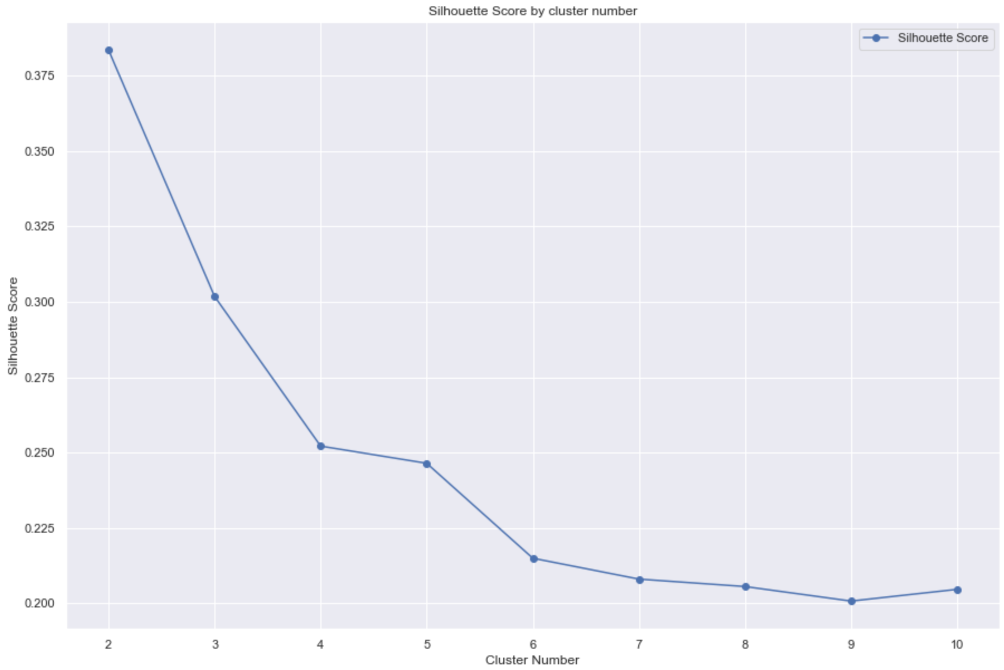


Figure 7: Average silhouette score of the clusters resulting from k-means-clustering for a given number of clusters k

Residuals from linear interpolations of scores for the 2 neighbouring k were calculated for k between 3 and 9 as described in the method section. The resulting graph is shown in Figure 8. Residuals are positive (meaning the silhouette score is better than expected from the interpolation) at 5 and 8 clusters with 5 being the global maximum and the residual at 8 being close to 0. This mathematically supports the visual identification of k=5 as positive outlier.

Overall, the average silhouette score method results in 2 and 5 as candidates for k; 2 from being the global maximum and 5 for being both a visual outlier and the global maximum of the residuals.

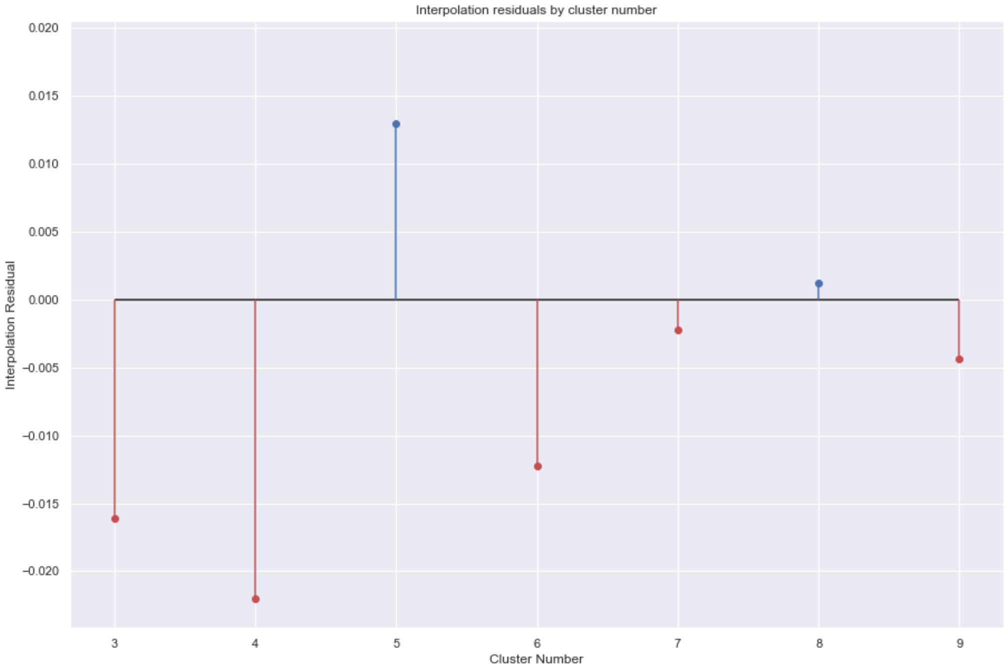


Figure 8: Residuals of observed average silhouette score of the clusters resulting from k-means-clustering for a given number of clusters k compared to interpolation. For calculation of the residuals, a linear interpolation is calculated between the scores of the neighbouring cluster numbers k-1 and k+1.

Clustering results for selected k

The elbow method and the average silhouette score method combined give 2, 3, 4, and 5 as candidates for the number of clusters (k) with 5 being the only candidate resulting from both methods. Graphs and reports for all k between 2 and 10 were created, but for sake of brevity only results for the candidate k and 6 as example of a non-candidate k are reported and discussed. The additional graphs and reports can be found in the appendix.

For 2 clusters (k=2), the average HAMD scores are shown in Figure 9. Both clusters show a monotonous decrease of average relative HAMD score with cluster 0 decreasing to approximately 0.32 and cluster 1 decreasing to approximately 0.77.

The corresponding silhouette plot is shown in Figure 10. There are 407 (53.2%) patients assigned to cluster 0 and 358 (46.8%) patients assigned to cluster 1. 7 (0.9%) of patients have a negative silhouette score, all of them in cluster 1 (1.9% of cluster 1).

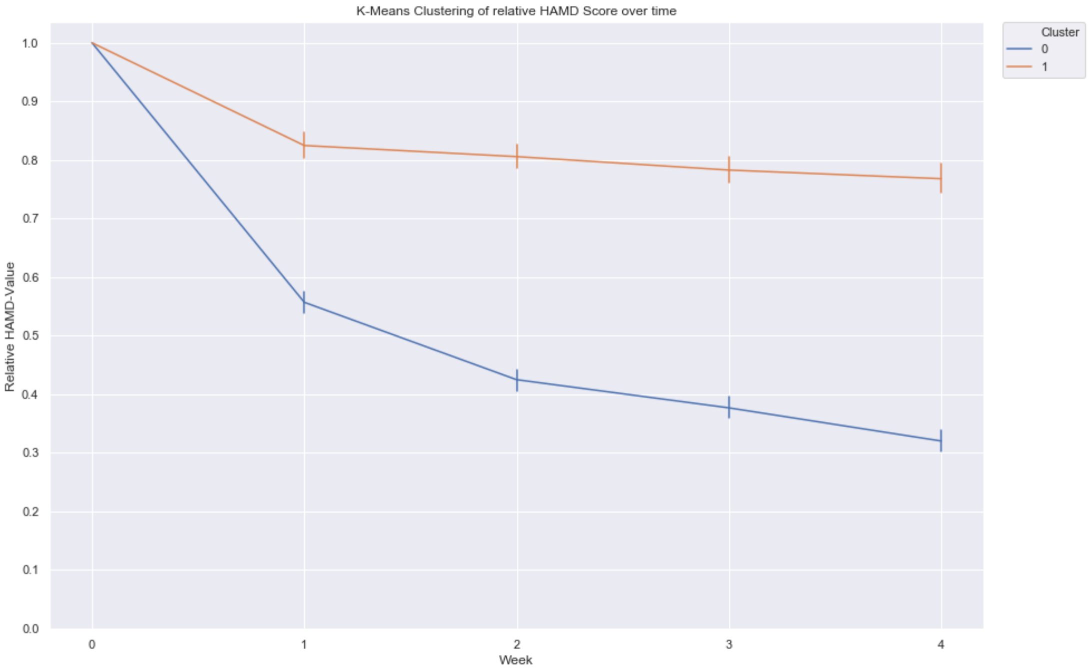


Figure 9: Average HAMD score by cluster resulting from k-means-clustering for 2 clusters. Error-Bars indicate the 95% confidence interval for the mean value.

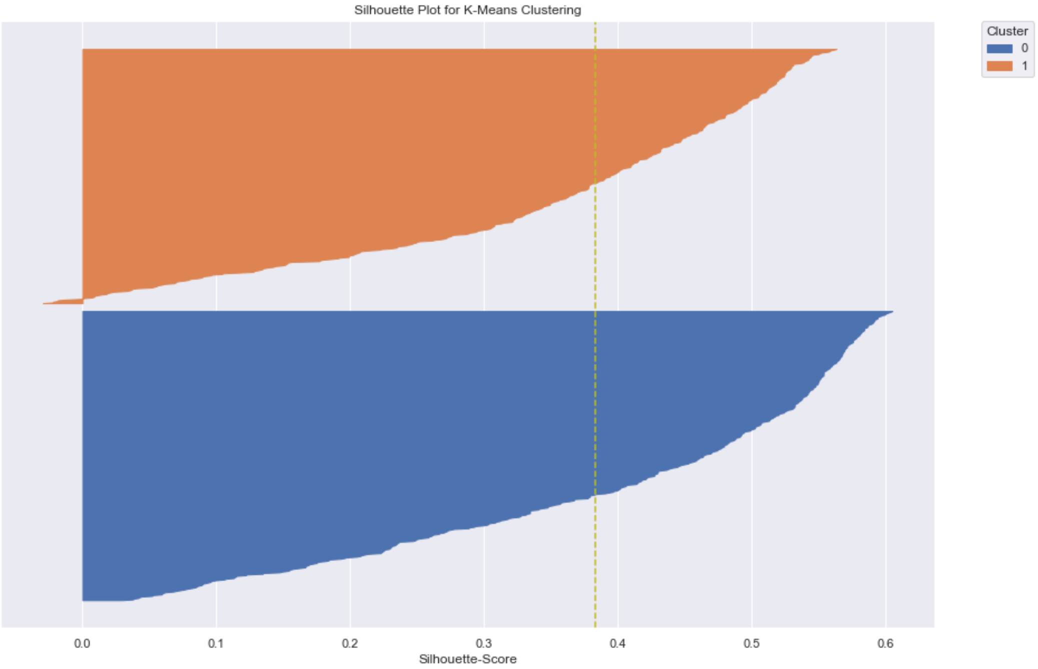


Figure 10: Silhouette plots for clusters resulting from k-means-clustering for 2 clusters. The average silhouette score is shown as a vertical line.

For 3 clusters (k=3), the average HAMD scores are shown in Figure 11. Both clusters 0 and 1 show a monotonous decrease of average relative HAMD score with cluster 0 decreasing to approximately 0.21 and cluster 1 decreasing to approximately 0.52. Cluster 2 shows a small decrease to approximately 0.89 in week 1 and stays at the same level after that.

The corresponding silhouette plot is shown in Figure 12. There are 210 (27.5%) patients assigned to cluster 0, 363 (47.5%) patients assigned to cluster 1 and 192 (25.1%) patients assigned to cluster 2. 14 (1.8%) of patients have a negative silhouette score, 8 of them in cluster 1 (2.2% of cluster 1) and 6 in cluster 2 (3.1% of cluster 2).

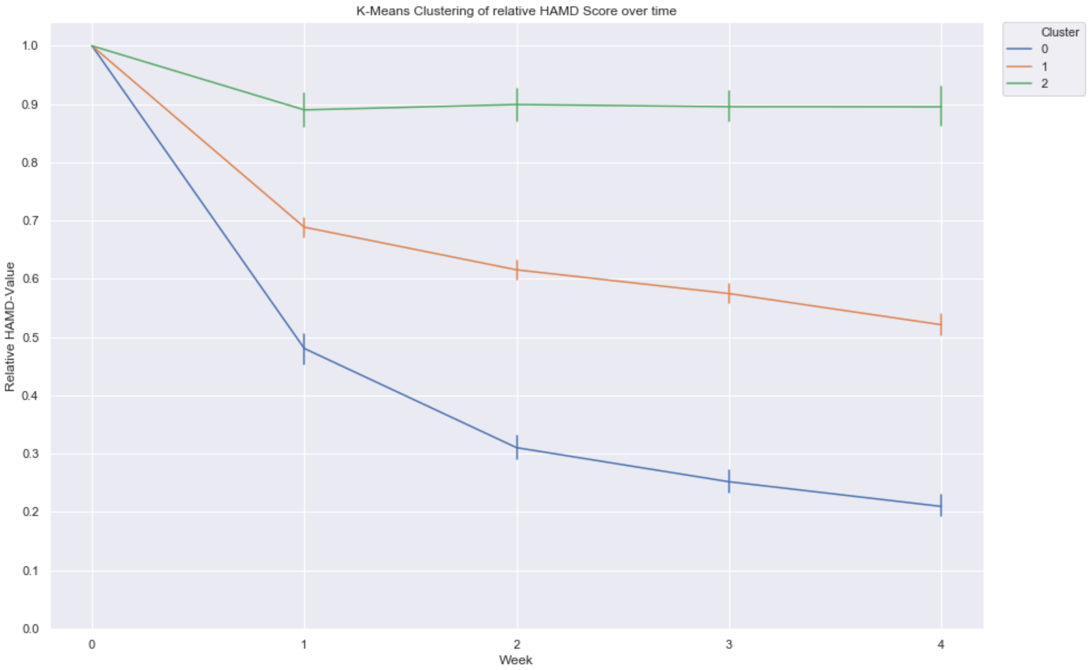


Figure 11: Average HAMD score by cluster resulting from k-means-clustering for 3 clusters. Error-Bars indicate the 95% confidence interval for the mean value.

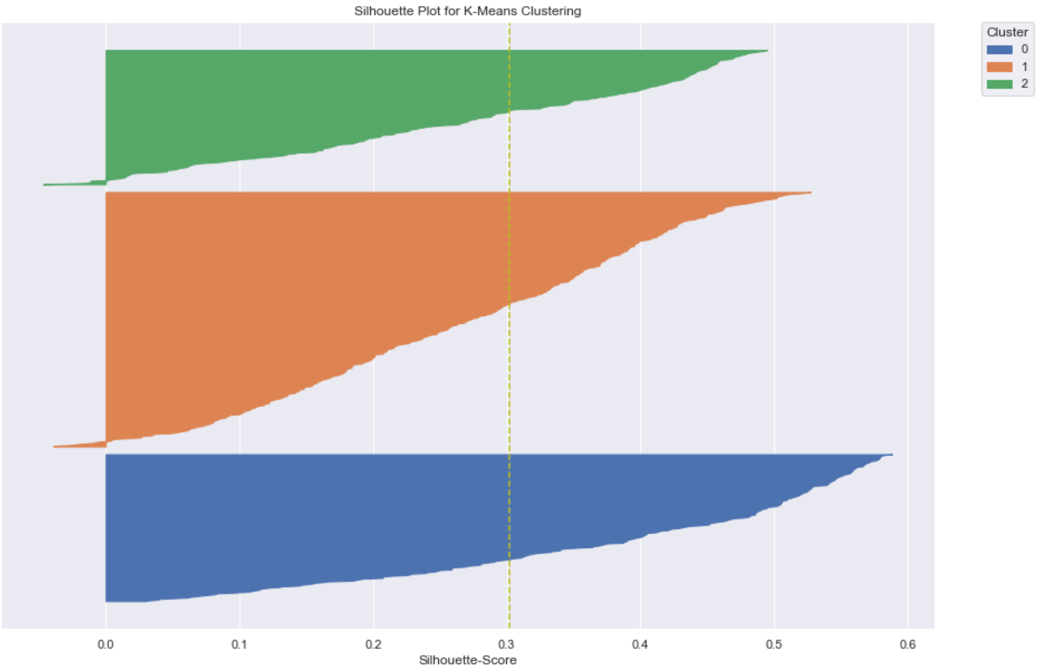


Figure 12: Silhouette plots for clusters resulting from k-means-clustering for 3 clusters. The average silhouette score is shown as a vertical line.

For 4 clusters (k=4), the average HAMD scores are shown in Figure 13. Clusters 0, 1 and 2 show a monotonous decrease of average relative HAMD score with cluster 0 decreasing to approximately 0.18, cluster 1 decreasing to approximately 0.42 and cluster 2 decreasing to approximately 0.67. Cluster 3 shows a small decrease to approximately 0.94 in week 1 and monotonously increases to approximately 1.03 until week 4.

The corresponding silhouette plot is shown in Figure 14. There are 155 (20.2%) patients assigned to cluster 0, 272 (35.6%) patients assigned to cluster 1, 237 (31.0%) patients assigned to cluster 2 and 101 (13.2%) patients assigned to cluster 3. 21 (2.7%) of patients have a negative silhouette score, 9 of them in cluster 1 (3.3% of cluster 1), 3 in cluster 2 (1.3% of cluster 2) and 9 in cluster 3 (8.9% of cluster 3).

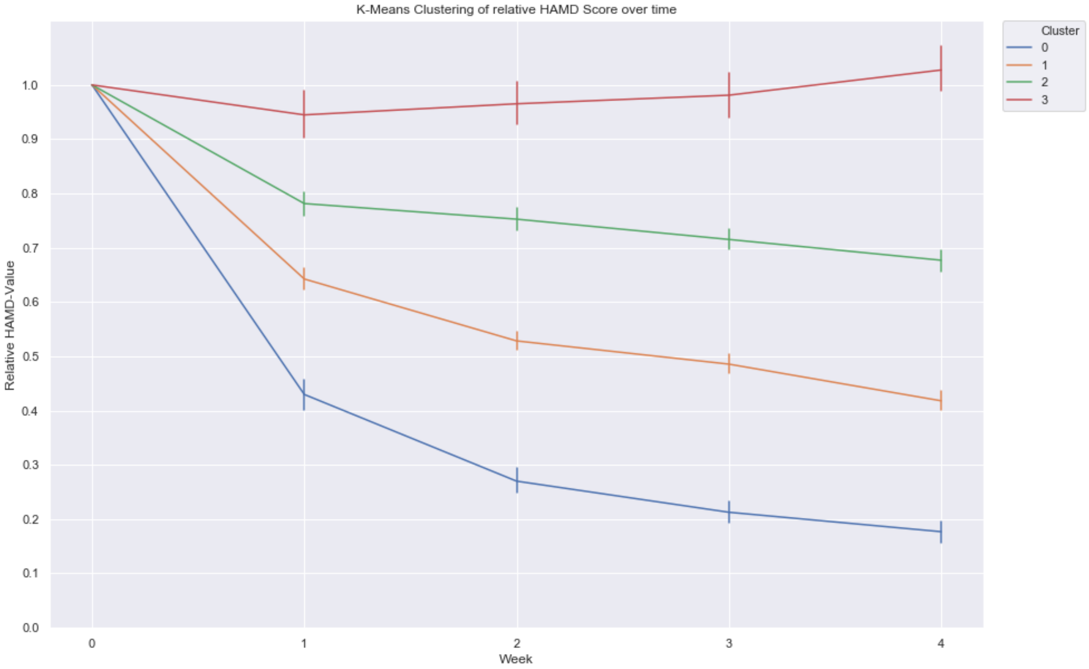


Figure 13: Average HAMD score by cluster resulting from k-means-clustering for 4 clusters. Error-Bars indicate the 95% confidence interval for the mean value

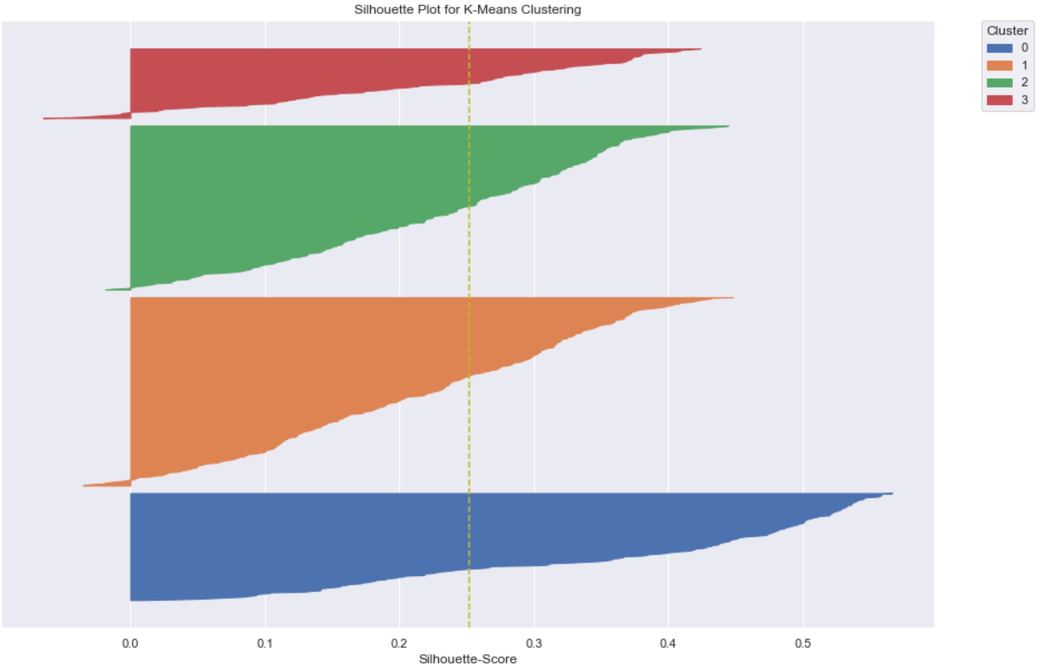


Figure 14: Silhouette plots for clusters resulting from k-means-clustering for 4 clusters. The average silhouette score is shown as a vertical line.

For 5 clusters (k=5), the average HAMD scores are shown in Figure 15. Cluster 0 shows a monotonous decrease to an average relative HAMD score of approximately 0.18 in week 4. Clusters 1 and 3 both show a similar initial decrease in week 1 (0.65 and 0.64 respectively) and diverge after that with cluster 1 decreasing monotonously to 0.39 and cluster 3 increasing to 0.75. A similar pattern is shown by clusters 2 and 4. Both show a similar initial decrease in week 1 (0.90 and 0.95 respectively) with cluster 2 decreasing monotonously to 0.57 and cluster 4 increasing monotonously to 1.03.

The corresponding silhouette plot is shown in Figure 16. There are 152 (19.9%) patients assigned to cluster 0, 235 (30.7%) patients assigned to cluster 1, 128 (16.7%) patients assigned to cluster 2, 155 (20.3%) patients assigned to cluster 3 and 95 (12.4%) patients assigned to cluster 4. 17 (2.2%) of patients have a negative silhouette score, 4 of them in cluster 1 (1.7% of cluster 1), 4 in cluster 2 (3.1% of cluster 2) and 9 in cluster 4 (9.5% of cluster 4).

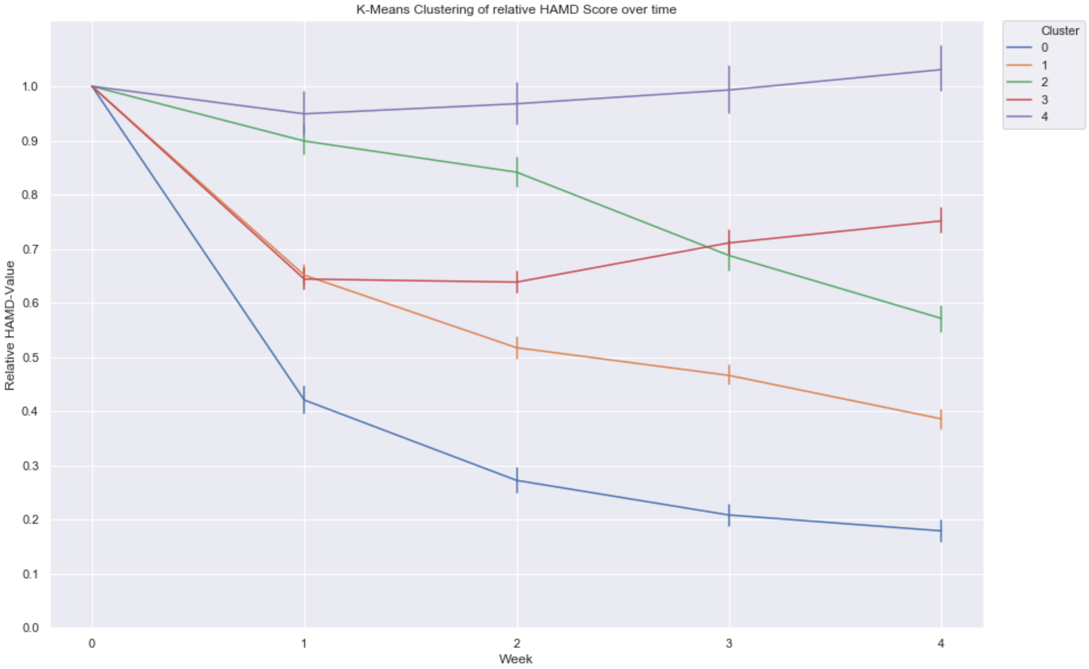


Figure 15: Average HAMD score by cluster resulting from k-means-clustering for 5 clusters. Error-Bars indicate the 95% confidence interval for the mean value

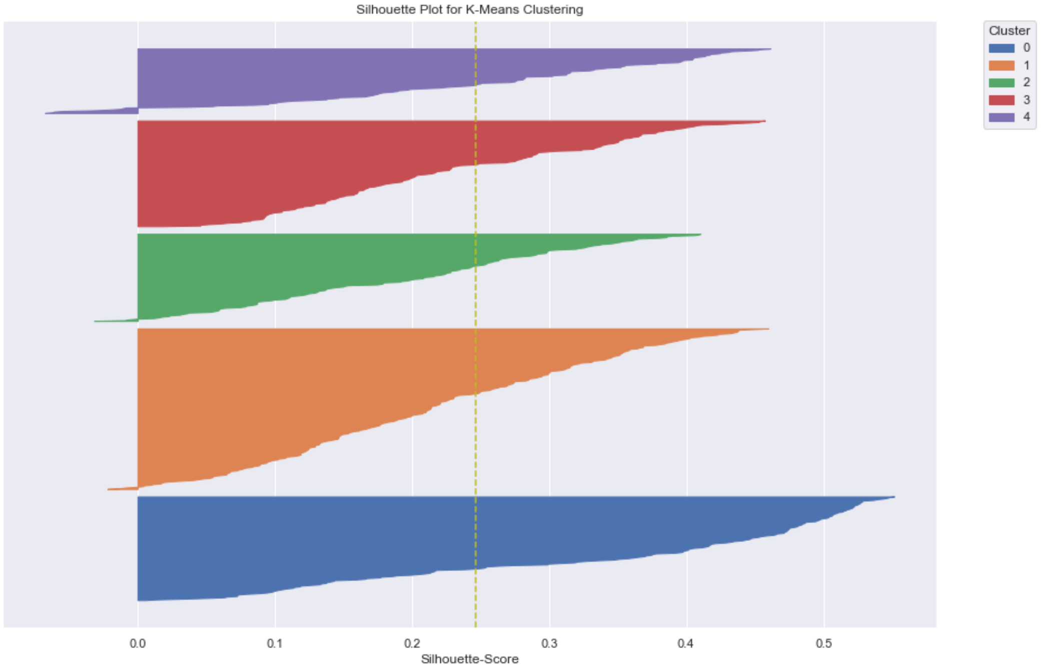


Figure 16: Silhouette plots for clusters resulting from k-means-clustering for 5 clusters. The average silhouette score is shown as a vertical line.

For 6 clusters (k=6), the average HAMD scores are shown in. Clusters 0, 1 and 3 show a monotonous decrease to an average relative HAMD score of approximately 0.15, 0.28 and 0.54 respectively in week 4. Cluster 2 shows a decrease to approximately 0.58 in week 1 and roughly stays at that level. Cluster 4 shows a decrease to approximately 0.72 in week 1 and an increase to 0.81 by week 4. Cluster 5 stays roughly at the initial level with a final average relative HAMD score of 1.04 in week 4.

The corresponding silhouette plot is shown in. There are 101 (13.2%) patients assigned to cluster 0, 159 (20.8%) patients assigned to cluster 1, 173 (22.6%) patients assigned to cluster 2, 126 (16.5%) patients assigned to cluster 3, 120 (15.7%) patients assigned to cluster 4 and 86 (11.2%) patients assigned to cluster 5. 36 (4.7%) of patients have a negative silhouette score, 9 of them in cluster 1 (5.7% of cluster 1), 1 in cluster 2 (0.6% of cluster 2), 16 in cluster 3 (12.7% of cluster 3) and 10 in cluster 5 (11.6% of cluster 5).

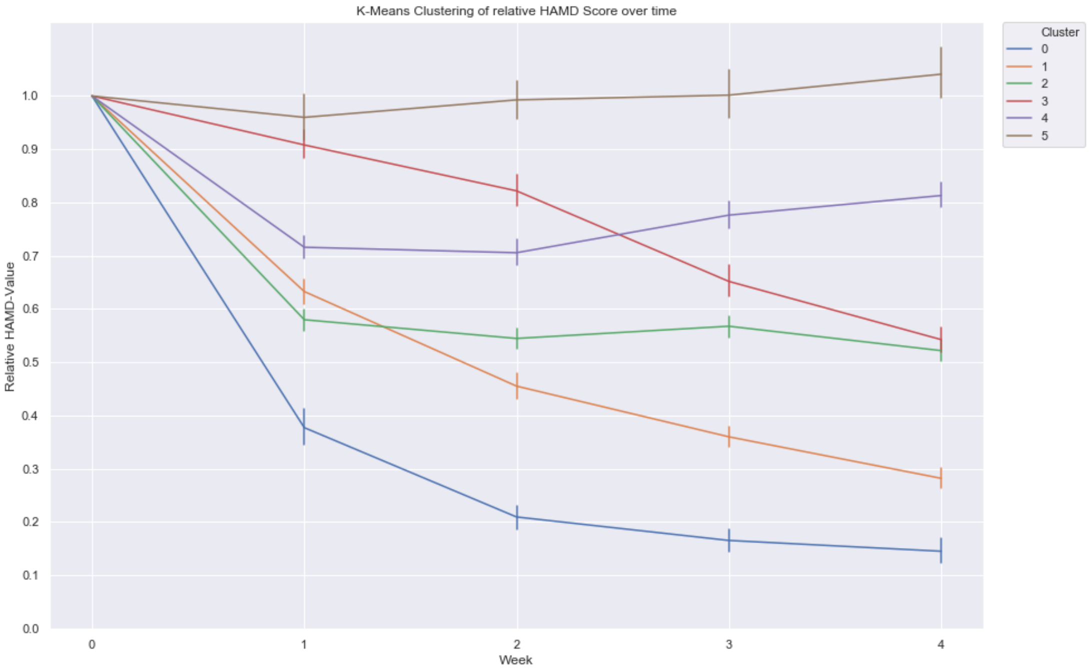


Figure 17: Average HAMD score by cluster resulting from k-means-clustering for 6 clusters. Error-Bars indicate the 95% confidence interval for the mean value

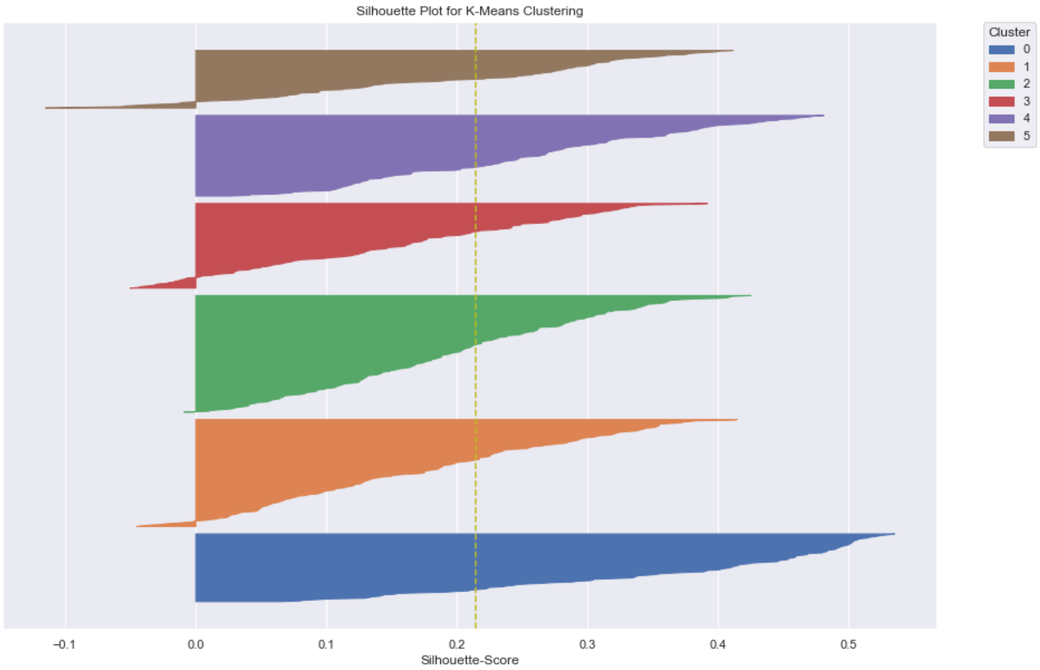


Figure 18: Silhouette plots for clusters resulting from k-means-clustering for 6 clusters. The average silhouette score is shown as a vertical line.

A summary of cluster sizes and observations with negative silhouette scores for k between 2 and 6 is also reported in Table 4 and Table 5 respectively.

N (%)	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
k=2	407 (53.2%)	358 (46.8%)	X	X	X	X
k=3	210 (27.5%)	363 (47.5%)	192 (25.0%)	X	X	X
k=4	155 (20.2%)	272 (35.6%)	237 (31.0%)	101 (13.2%)	X	X
k=5	152 (19.9%)	235 (30.7%)	128 (16.7%)	155 (20.3%)	95 (12.4%)	X
k=6	101 (13.2%)	159 (20.8%)	173 (22.6%)	126 (16.5%)	120 (15.7%)	86 (11.2%)

Table 4: Cluster size structure for clusters resulting from k-means-clustering for a given number of clusters k. Sums differing from 100% are due to rounding error.

Neg. Silh. Score (%)	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Total
k=2	0 (0%)	7 (2.0%)	X	X	X	X	7 (0.9%)
k=3	0 (0%)	8 (2.2%)	6 (3.1%)	X	X	X	14 (1.8%)
k=4	0 (0%)	9 (3.3%)	3 (1.3%)	9 (8.9%)	X	X	21 (2.7%)
k=5	0 (0%)	4 (1.7%)	4 (3.1%)	0 (0.0%)	9 (9.4%)	X	17 (2.2%)
k=6	0 (0%)	9 (5.7%)	1 (0.6%)	16 (12.7%)	0 (0.0%)	10 (11.6%)	36 (4.7%)

Table 5: Patients with negative silhouette scores for clusters resulting from k-means-clustering for a given number of clusters k. Percentages for the individual clusters are relative to n for that cluster, percentages for total are relative to total n.

Validation results

Variance explained and average silhouette scores were calculated in the validation dataset with trained algorithms for k between 2 and 10 as described in the methods section. An overview of results is shown in Figure 19. Both validation variance explained, and silhouette score show high similarity to the corresponding training graphs. The validation silhouette score for all candidate k (2, 3, 4 and 5) is higher than the corresponding training score. The validation variance explained is smaller than its training counterpart for k=2 and higher for the other candidate k.

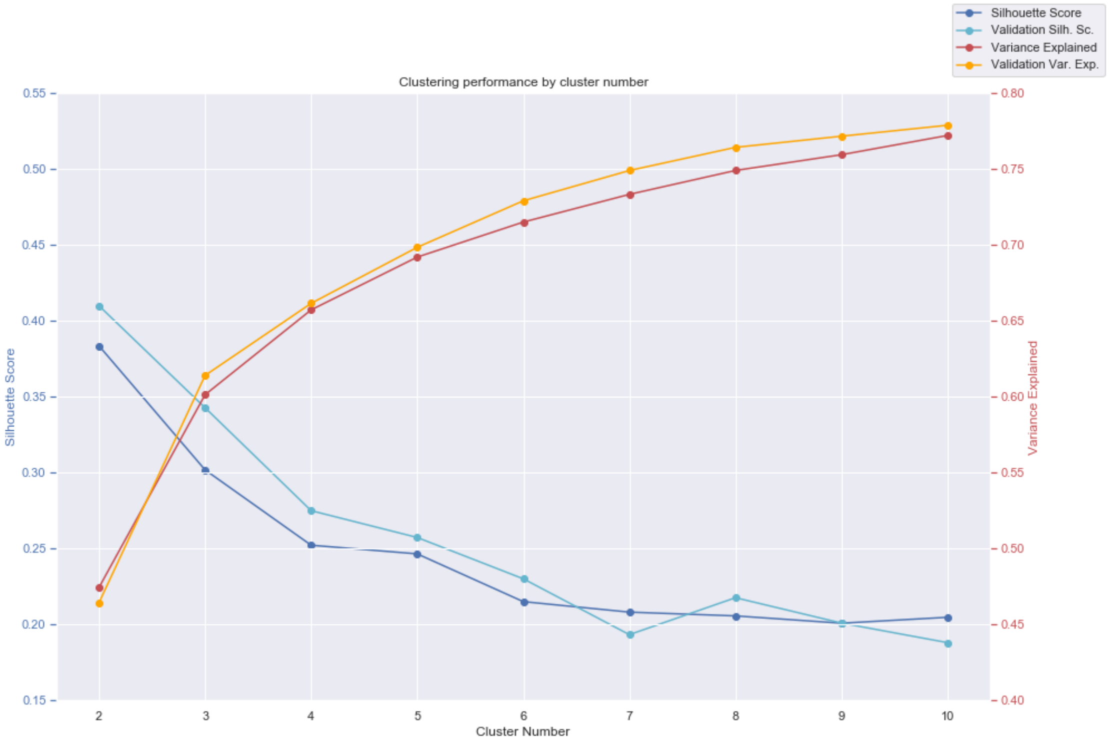


Figure 19: Variance explained and Silhouette score for clusters resulting from k-means-clustering for a given number of clusters k. Metrics are shown for training and validation dataset in comparison.

Structure of cluster sizes and observations with negative silhouette scores are reported in Table 6 and Table 7 respectively. For sake of brevity, only candidate k and 6 as example of a non-candidate k are reported. For sake of additional brevity, graphs of mean relative HAMD scores and silhouette plots aren’t shown here. As far as they are referenced in the discussion section, they are shown there. Full graphs can be found in the supplementary material.

N (%)	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
k=2	628 (61.1%)	400 (38.9%)	X	X	X	X
k=3	356 (34.6%)	448 (43.6%)	224 (21.8%)	X	X	X
k=4	255 (24.8%)	400 (38.9%)	249 (24.2%)	124 (12.1%)	X	X
k=5	236 (23.0%)	368 (35.8%)	203 (19.7%)	101 (9.8%)	120 (11.7%)	X
k=6	148 (14.4%)	305 (29.7%)	147 (14.3%)	213 (20.7%)	101 (9.8%)	114 (11.1%)

Table 6: Cluster size structure for clusters resulting from applying trained k-means-clustering for a given number of clusters k to the validation dataset. Sums differing from 100% are due to rounding error.

Neg. Silh. Score (%)	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Total
k=2	0 (0%)	19 (4.7%)	X	X	X	X	19 (1.8%)
k=3	1 (0.3%)	3 (0.7%)	12 (5.4%)	X	X	X	16 (1.6%)
k=4	1 (0.4%)	17 (4.3%)	10 (4.0%)	3 (2.4%)	X	X	31 (3.0%)
k=5	1 (0.4%)	21 (5.7%)	19 (9.4%)	1 (1.0%)	6 (5%)	X	48 (4.7%)
k=6	2 (1.4%)	57 (18.7%)	0 (0%)	39 (18.3%)	4 (4.0%)	8 (7.0%)	110 (10.7%)

Table 7: Patients with negative silhouette scores for clusters resulting from applying trained k-means-clustering for a given number of clusters k to the validation dataset. Percentages for the individual clusters are relative to n for that cluster, percentages for total are relative to total n.

Descriptive Reports

As described in the method section, reports of between-cluster differences (and their statistical significance) in sociodemographic and clinical variables were created for all k between 2 and 10. For reasons of brevity, these reports are not shown here. As far as they are referenced in the discussion section, they are shown there.

Discussion

Selecting the number of clusters

Different candidates for a given number of clusters (k) have been identified as described in the results section for experiment 1. The elbow method gave 3, 4 and 5 as possible candidates and the average silhouette score gave 2 as candidate from the global maximum and 5 as candidate for being a visual outlier and the maximum of the residuals from interpolation.

In order to generate hypotheses for clinical decision making (as explained on page 16) based on cluster structure, choosing one or more mathematically valid, generalizable and clinically interpretable cluster structures is key (See page 25). Mathematical validity was operationalized with the goodness of fit data, generalization was assessed with the external validation dataset and clinical interpretability will be discussed in combination with the other factors in this subsection.

The candidates for k are discussed individually, in comparison to other candidates and compared to k=6 as example for a non-candidate below.

k=2 is strongly supported by the global maximum for the silhouette score. The low percentage of patients with a negative silhouette score in both training (0.9%) and validation (1.8%) datasets adds further support. The silhouette plot shows similar sized, smooth, convex silhouettes in both training (Figure 10) and validation dataset (Figure 20).

But while the supporting data is strong, interpretational usefulness of 2 clusters is questionable. The mean relative HAMD scores for cluster 0 go below the level of 0.5 which is commonly used as definition of response, whereas cluster 1 doesn't reach that level. Therefore, interpreting the 2 clusters as responders and non-responders seems obvious. Following this interpretation, the question needs to be asked, whether the clustering approach gives any advantage over the traditional definition of response. Since the clustering aligns with the common clinical definitions without adding any additional information, there is no information from which to generate new interventional hypotheses. Argued inversely, the data supporting 2 clusters are supporting the traditional response definition.

In summary, k=2 is strongly supported by goodness of fit data but is interpretationally non-superior to the traditional definition of response. Occam's Razor also applies, since the traditional definition is far simpler than the clustering approach. For those reasons, k=2 is a poor candidate for hypothesis generation.

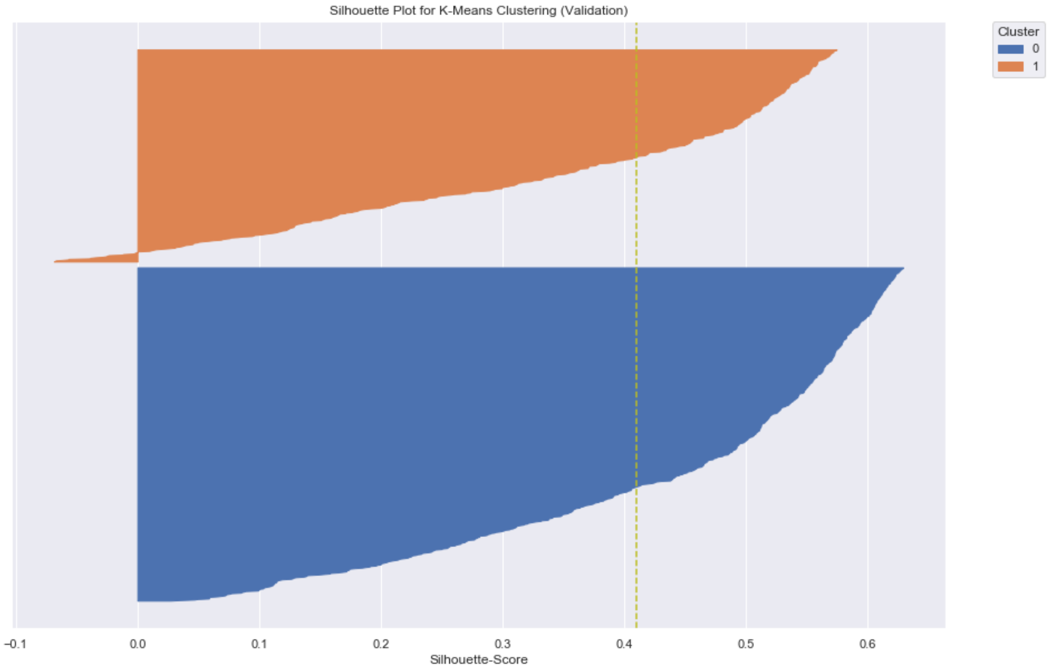


Figure 20: Silhouette plots for clusters resulting from applying trained k-means-clustering for 2 clusters to the validation dataset. The average silhouette score is shown as a vertical line.

The supporting goodness of fit data is far weaker for k=3. While it is a candidate given by the elbow method, the average silhouette score method shows neither a local maximum nor a positive residual to linear interpolation. The silhouette plot (Figure 12) is both less smooth and less convex that for k=2. On the hand of positive evidence, the number of patients with negative silhouette scores is low and similar for both training (1.8%) and validation (1.6%).

Similar arguments are found for k=4. As another candidate from the elbow method, the average silhouette method is non-supportive. The number of patients with negative silhouette scores in training is higher than for both k=3 and k=5 (2.7% vs. 1.8% and 2.2% respectively) which weakens the goodness of fit data further.

The remaining candidate k=5 has a strong case from being both a candidate from the elbow method and the silhouette score method. In the latter, it is both a visual outlier and the global maximum of linear interpolation residuals. The silhouette plots for training (Figure 16) and validation both show smooth outlines with minimal concavity. The number of patients with negative silhouette scores is lower than for both k=4 and k=6 (2.2% vs. 2.7% and 4.7% respectively) in training with an increase to 4.7% in the validation set. The difference in the training set can be considered evidence of good fit compared to the neighbours. The increase in the validation dataset can be considered evidence of overfitting, meaning the algorithm captured some of the noise in the training data which results in comparatively poor results in the validation set. Overfitting can occur in most machine learning algorithms and commonly increases with model complexity, so an increase of overfitting can be expected with growing k without necessarily being indicative of poor choice of k. In line with that expectation, there is an increase of patients with negative silhouette score in the validation set for all k >= 4. So, while the occurrence of overfitting mildly weakens the goodness of fit data for k=5, the same arguments can be made for k=4 and k=6.

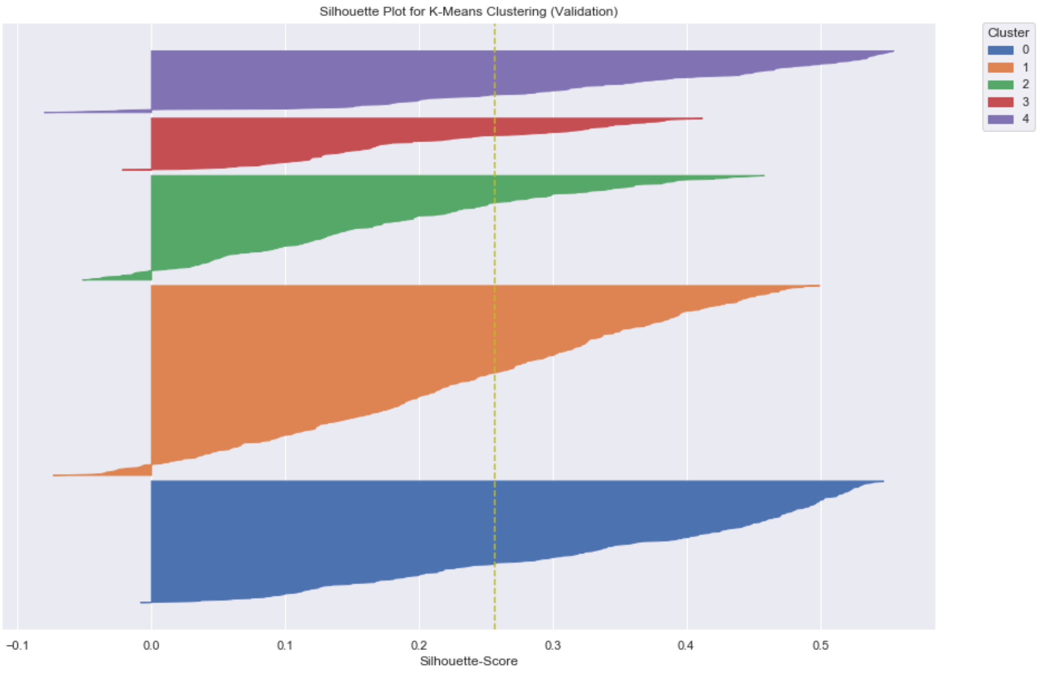


Figure 21: Silhouette plots for clusters resulting from applying trained k-means-clustering for 5 clusters to the validation dataset. The average silhouette score is shown as a vertical line.

k=6 as example for a non-candidate shows comparatively weak goodness of fit evidence as expected. Both elbow-method and average silhouette score method don’t give it as candidate, the number of patients with negative silhouette score is high in the training set (4.7%) with a large increase in the validation set (10.7%) suggesting overfitting as described above. The silhouette plot (Figure 18) shows some large negative outliers to the overall smooth and only mildly concave shapes.

To summarize the goodness of fit data, of the different candidates for k > 2, there is comparable evidence for k=3 and k=4. k=5 has stronger evidence in comparison to both. In order to be useful for hypothesis generation, the clustering should both show good fit (indicating mathematical validity for the training set and generalization on the validation set) and be useful for clinical and research interpretation. The graphs of mean relative HAMD scores for k=3 and k=4 (Figure 11 and Figure 13 respectively) show different degrees of response that run mostly parallel. While this might allow for a more fine-grained approach in clinical decision making, the interpretational case for k=5 is much stronger. In the corresponding graph (Figure 15) we see 3 initial groups of patients: Those showing early response (cluster 0), those showing early improvement (clusters 1 and 3) and those showing no early improvement (clusters 2 and 4). Of the initial non-improvers, cluster 2 goes on to show a delayed improvement and cluster 4 continues to show non-improvement. Of the early improvers, cluster 1 continues to show response and cluster 3 shows non-response. Interpreting the clusters in this 2-level-approach seems obvious, since the traditional response definitions and early improvement definitions (>50% and >20% decrease from baseline) can be applied on the level of the cluster mean values. This gives a strong interpretational case due to the strong alignment to the traditional definitions and thus, previous evidence, while at the same time adding additional time course information. Due to the interpretational analysis supporting k=5 and goodness of fit data being comparatively supportive of k=5 as well, choosing the resulting clusters from k=5 as clusters for hypothesis generation is in the authors opinion justified, whereas no hypotheses should be derived from the cluster structures for k=3 or k=4.

Since only k=5 was selected for further hypothesis generation in this discussion subsection, from this point onward all mentions of “cluster structure” or “clusters” will reference the structure for k=5 unless otherwise indicated. For ease of reference, the clusters for k=5 are given free-text names based on the interpretation described in detail above with the original cluster numbers given in brackets from this point forward. Free text names are shown in the legend of Figure 22. The group of clusters 1 and 3 will be referred to as “Early Improvement” and the group of Clusters 2 and 4 as “Early Non-Improvement”.

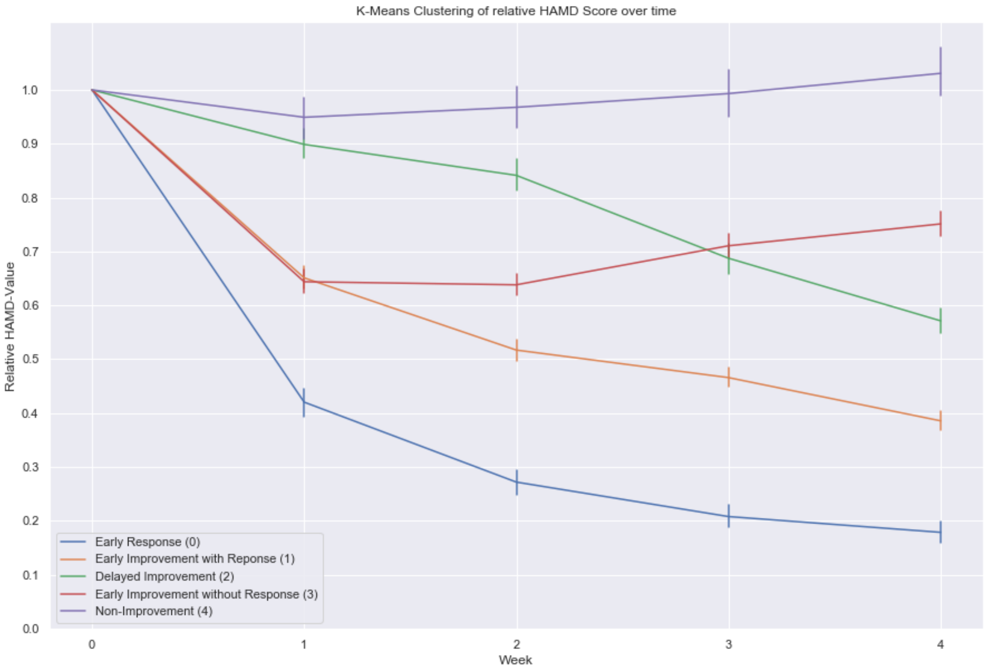


Figure 22: Average HAMD score by cluster resulting from k-means-clustering for 5 clusters. Error-Bars indicate the 95% confidence interval for the mean value. Free text names for interpretation are shown in the legend.

Comparison to previous cluster structures

If the clusters for k=5 are compared to the first few weeks of the latent classes identified by Uher et al. (2011), there are several parallels [10]. In order to compare the groups side by side, data for the first 4 weeks was extracted from the graph shown in Figure 2 via WebPlotDigitizer, a Computer Tool to extract graph data [47]. The resulting comparison graph is shown in Figure 23. The Non-Improvement (4) cluster from this thesis is comparable to a combination of classes 1, 3 and 4. Additionally, the Delayed Improvement cluster (2) is comparable to a combination of classes 2, 5, 7 and 8. The Early Response cluster (0) is roughly comparable to class 9 and the Early Improvement with Response (1) cluster is comparable to class 6. Only the Early Improvement without Response (3) cluster has no parallel classes. This could be the results of this response pattern being added into one or more the different classes that run parallel to the Delayed Improvement (2) or Early improvement with Response (1) cluster.

These parallels might have significant implications as to clinical usefulness of the cluster structure developed in this thesis. For example, the Delayed Improvement cluster (2) runs roughly parallel to both classes 7 and 8, which are classes with the second and third lowest final relative HAMD score. This could imply that a subset of patients from the Delayed Improvement cluster (2) would benefit from taking the same antidepressant medication over a period longer than 4 weeks, even though the cluster average does not reach the traditional response criterium at week 4.

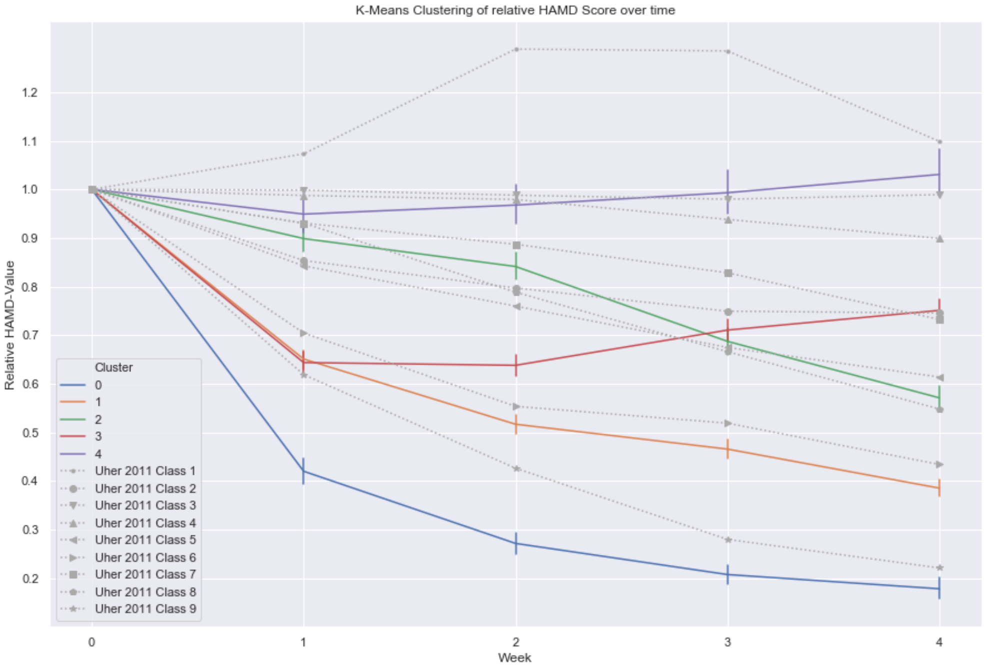


Figure 23: Average HAMD score by cluster resulting from k-means-clustering for 5 clusters. Error-Bars indicate the 95% confidence interval for the mean value. For comparison, Classes from [10] have been drawn in the graphic.

A similar comparison to the treatment classes suggested by Paul et al. (2019) isn’t immediately possible due to the differing underlying assumptions in their methodology [33]. By using logarithmically transformed HAMD sum scores and fitting a (linear) mixed effects model, the authors are essentially fitting an optimal number of logarithmic functions. This constraint to finding logarithmic clusters is not present in the current study or in Uher et al. (2009) [10]. The cluster structure resulting from this methodology (as shown in Figure 3) captures this underlying assumption by showing average slopes “fanning” a large portion of the logarithmic space. While this might not necessarily constitute a weakness of the methodology, as the validation as to the clusters predictive value by Paul et al. (2019) shows, hypothesis inferred from clustering without a “shape”-constraint like in the current thesis does naturally differ widely based on methodology alone [33].

It should further be noted, that both the aforementioned structures of treatment response clusters examined a longer timeframe of the treatment course. By limiting the clustering effort to the first four weeks of antidepressant response, the clustering proposed here emphasises the early timeframe of treatment, in order to gather predictive information that might not be captured by the longer duration examination. On the other hand, the longer duration clustering will likely be more useful for assessment of total treatment courses.

Between-Cluster Differences

Having chosen k=5 for hypothesis generation results in the aforementioned clusters of early response. To facilitate earlier clinical decision making, classifying is only the first step. Prediction of a patient's cluster earlier than week 4 is one possibility to outperform current clinical guidelines (see page 16). To make prediction feasible, there should be differences between the patients belonging to different clusters. Checking for these differences in the descriptive reports that have been generated can serve as sanity and feasibility check before attempting more complicated predictive algorithms. Table 8 and Table 9 give an overview of the descriptive variables for the different clusters.

Variable: N (%)	Early Response (0)	Early Improvement with Response (1)	Delayed Improvement (2)	Early Improvement without Response (3)	Non-Improvement (4)
Gender (Chi²-p: 0.798)					
Male	69 (45.4%)	105 (44.7%)	53 (41.4%)	62 (40.0%)	38 (40.0%)
Female	83 (54.6%)	130 (55.3%)	75 (58.6%)	93 (60.0%)	57 (60.0%)
Ethnic Group (Chi²-p: 0.888)					
European	145 (95.4%)	230 (97.9%)	123 (96.1%)	149 (96.1%)	92 (96.8%)
Asian	2 (1.3%)	0 (0%)	1 (0.8%)	3 (1.9%)	1 (1.1%)
African	2 (1.3%)	2 (0.9%)	1 (0.8%)	1 (0.6%)	0 (0%)
Other	3 (2.0%)	3 (1.3%)	3 (2.3%)	2 (1.3%)	2 (2.1%)
Highest School Degree (Chi²-p: 0.928)					
None	3 (2.0%)	3 (1.3%)	1 (0.8%)	0 (0%)	2 (2.1%)
Lower secondary school	48 (31.6%)	76 (32.3%)	42 (32.8%)	40 (25.8%)	29 (30.5%)
Intermediate secondary school	49 (32.2%)	68 (28.9%)	39 (30.5%)	51 (32.9%)	30 (31.6%)
Advanced technical certificate	17 (11.2%)	29 (12.3%)	16 (12.5%)	23 (14.8%)	8 (8.4%)

Variable: N (%)	Early Response (0)	Early Improvement with Response (1)	Delayed Improvement (2)	Early Improvement without Response (3)	Non-Improvement (4)
Upper secondary school	33 (21.7%)	57 (24.3%)	30 (23.4%)	39 (25.2%)	26 (27.4%)
Other	2 (1.3%)	2 (0.9%)	0 (0%)	2 (1.3%)	0 (0%)
Highest Vocational Degree (Chi^2-p: 0.332)					
None	19 (12.5%)	23 (9.8%)	16 (12.5%)	30 (19.4%)	14 (14.7%)
Apprenticeship	61 (59.9%)	144 (61.3%)	76 (59.4%)	79 (51.0%)	56 (58.9%)
Master	6 (3.9%)	6 (2.6%)	1 (0.8%)	6 (3.9%)	1 (1.1%)
University, College of higher Education	27 (17.8%)	50 (21.3%)	31 (24.2%)	37 (23.9%)	18 (18.9%)
Vocational College	7 (4.6%)	10 (4.3%)	4 (3.1%)	3 (1.9%)	4 (4.2%)
Other	2 (1.3%)	2 (0.9%)	0 (0%)	0 (0%)	2 (2.1%)
Recurrent MDD (Chi^2-p: 0.036*)					
First Episode	63 (41.4%)	85 (36.2%)	36 (28.1%)	52 (33.5%)	23 (24.2%)
Previous Episodes	89 (58.6%)	150 (63.8%)	92 (71.9%)	103 (66.5%)	72 (75.8%)
Age of MDD Onset (Chi^2-p: 0.002**)					
Early Onset (before Age 21)	29 (19.1%)	46 (19.6%)	42 (32.8%)	49 (31.6%)	25 (26.3%)
Middle Onset (Ages 21 to 44)	89 (58.6%)	126 (53.6%)	67 (52.3%)	75 (48.4%)	59 (62.1%)
Late Onset (After Age 45)	34 (22.4%)	63 (26.8%)	19 (14.8%)	30 (19.4%)	11 (11.6%)

*Table 8: Categorical demographic and clinical descriptors by Cluster from k-means-clustering for 5 Clusters. Differences in absolute sums are due to missing values, differences in percentage sums are due to rounding error. Percentage sums are referring to the ration within each cluster. Chi-Squared-Test results are indicated by p value. *= p< 0.05, **p<0.01, *** p<0.001.*

Variable: Mean (95% CI)	Early Response (0)	Early Improvement with Response (1)	Delayed Improvement (2)	Early Improvement without Response (3)	Non-Improvement (4)	ANOVA-p
Age	41.9 (95% CI 40.2 to 43.7)	41.3 (95% CI 39.9 to 42.8)	38.3 (95% CI 36.2 to 40.5)	40.6 (95% CI 38.7 to 42.5)	39.7 (95% CI 37.3 to 42.2)	p=0.874
Years of Education	13.9 (95% CI 13.4 to 14.4)	14.2 (95% CI 13.8 to 14.6)	13.9 (95% CI 13.2 to 14.5)	14.0 (95% CI 13.4 to 14.5)	13.5 (95% CI 12.8 to 14.3)	p=0.677
Age at MDD onset	34.5 (95% CI 32.6 to 36.5)	34.0 (95% CI 32.4 to 35.6)	29.8 (95% CI 27.6 to 31.9)	31.2 (95% CI 29.2 to 33.2)	29.6 (95% CI 27.4 to 31.9)	p<0.001 ***
Number of previous MDD episodes	2.29 (95% CI 1.74 to 2.84)	2.31 (95% CI 1.87 to 2.75)	2.35 (95% CI 1.79 to 2.92)	2.90 (95% CI 1.97 to 3.82)	3.93 (95% CI 2.53 to 5.33)	p=0.031 *
Length of current episode (days)	24.3 (95% CI 19.4 to 29.3)	29.6 (95% CI 24.2 to 35)	39.2 (95% CI 23.8 to 54.6)	36.4 (95% CI 27.7 to 45.1)	31.8 (95% CI 22.7 to 40.9)	p=0.147
HAMD score at Baseline	22.9 (95% CI 22.2 to 23.5)	23.5 (95% CI 22.9 to 24.1)	21.7 (95% CI 20.9 to 22.5)	24.2 (95% CI 23.5 to 24.9)	21.0 (95% CI 20 to 22)	p<0.001 ***

*Table 9: Continuous demographic and clinical descriptors by Cluster from k-means-clustering for 5 Clusters. 95% CI: 95 % confidence interval for the mean. One-Way ANOVA results are indicated by p value. *= p< 0.05, **p<0.01, *** p<0.001.*

Statistically significant differences can be observed for the categorical variables Recurrence of MDD and MDD onset group and for the continuous variables Age of MDD onset, Number of previous episodes and HAMD score at baseline. The categorical variables correspond to recoding of the two first mentioned continuous variables, so significant differences in both corresponding variables is expected.

Existence of these baseline differences falls in line with expected differences from the previous literature. Overall symptom severity was a significant predictor of later treatment response in Chekroud et al. (2016), Rush et al. (2008) and Nierenberg et al. (2000) [31] [28] [22]. The theoretical importance of age of onset as significant predictor is emphasised by the existence of targeted studies like Kozel et al. (2008) [29]. Though that study didn't find a significant effect in its dataset, the underlying theoretical importance stands. The number of previous MDD episodes was one of the significant predictors in Chekroud et al. (2016) [31]. These parallels of between cluster differences and predictors of response from previous literature gives the cluster structure some face validity as to its possible usefulness and/or predictive value. At the same time, it's of interest to note, that common demographic response predictors (employment, ethnicity) don't show significant differences in the cluster structure. This could well be the result of insufficient statistical power (due to very small n for some groups), so the implications from this should not be overvalued.

Comparison with other cluster algorithm families

K-means-clustering enforces hard clustering. This means, every patient was assigned to a single cluster. While this makes the analysis easier, it doesn't adequately capture patients that might fall "between clusters" or outside of the cluster prototypes. This limitation was partially quantified in the result section, when patients with negative silhouette scores were reported. These patients are outliers from the cluster they were assigned to, despite it being the best fit. This algorithmic property resulting from the choice of clustering methodology has important implications for utilizing the clusters as tools for future scientific research or clinical decision making.

For scientific inquiry into differences between different pattern of early treatment response using a clustering approach like in this experiment, results might benefit from not treating all patients assigned to a cluster the same. Selections need to be made, how to deal with differing typicality of the patients, for example as measured by the silhouette score. Whether or not edge-cases or outliers of the cluster assignment should be in- or excluded for analysis or whether results might benefit from a typicality-weighted approach depends on the research context, but these questions should be kept in mind when opting for use of clustering methodology and the resulting tools.

Other clustering algorithms could be utilized as an alternative. Examples include so called "soft" or "fuzzy" clustering algorithms, that allow objects to be in more than one cluster at a time, often with some degree of membership or probability measure for each cluster. An example of these algorithms is the C-means-algorithm [48] that's closely related to the K-means-algorithm used here with the difference, that an object can belong to all groups with different membership grades between 0 and 1. Fuzzy clustering could solve issues with identifying outliers or directly provide weights for a prototypicality-weighted approach.

As another alternative, clustering algorithms based on Bayesian statistics might be used. The main strength of these algorithms is the incorporation of uncertainty and usage of the entire space of possible cluster-separations for probability estimates following Bayes rule (see [49] for detailed background a modern version of a Bayesian clustering algorithm). The resulting cluster structure is as statistical model and can be analysed as such as opposed to merely being a heuristic, like the k-means-algorithm used here. This specifically also allows to test the cluster structure versus the null hypothesis, that there are no clusters within the group. This is a critical – and commonly neglected – step if the suggested cluster structure is proposed to be the result of an underlying mechanism, as opposed to merely being a useful and mathematically valid classification of the parameter space (the latter being the case in this thesis).

Lastly, hierarchical clustering algorithms can be employed. This group of algorithms gives multiple degrees of separation. On the first level, all objects are treated as individual. The second level treats each set of the 2 most proximal objects as a cluster. In subsequent levels, objects are added to clusters or clusters are added to each other based on proximity as defined based on some proximity measure. The different levels are commonly visualised as dendrogram (see [50] for a more detailed introduction into hierarchical clustering). The main advantage of this approach is that relationships between different level clusters can be examined heuristically, which allows for easy interpretation. As a weakness, hierarchical clustering algorithms are often sensitive to small changes or noise in the dataset. There is additional difficulty in determining the number of clusters, since there often is no obvious level on which to “cut” the decision dendrogram.

The main advantage of the k-means-algorithm over the alternative clustering approaches presented here lies in model simplicity. Simpler models not only result in lower computational cost but also often make generalization of results easier following Occam’s razor. Scientist employing similar methodology should be aware of the different algorithmic families and should chose appropriately for the individual research question.

Experiment 2 – Prediction of clinical outcomes

Summary of Experiment 2

Random forest classifiers with differing variable sets (variables based on HAMD, IDS, both and a set additionally including demographic and clinical history variables) were trained to predict both traditional response and traditional remission criteria at week 4. Prediction performance was quantified using accuracy, sensitivity and specificity and both Chi²-Test and calculation of effect size (estimated by Cramer´s V) were performed. Feature importance scores were calculated for the different variable sets.

Predictive performance was comparable across models from all variable sets. When including only variables from baseline, performance was comparable to the zero-information rate. When variables from any later timepoints were included, prediction was possible significantly over random and both the 20% and 30% early improvement criteria, were outperformed. Prediction accuracy for later response increased monotonously with inclusion of timepoints after baseline with the largest increase between baseline and week 1, suggesting an early onset of antidepressant action within that timeframe.

The most important features for prediction were HAMD and IDS sum scores as well as principal components of their items and previously established demographic and clinical variables. Inclusion of these variables into a predictive dataset for clustering prediction was discussed.

Aim

The cluster structure identified in Experiment 1 provides a classification of early response patterns over time. In order to facilitate decision making before the 4-week timepoint based on that 5-cluster structure, a predictive model should be developed. In order to develop a model with maximum prediction accuracy, feature engineering is required. Feature engineering refers to a combination of transformation and (subsequent) selection of variables in order to create a dataset with the largest possible predictive value while simultaneously filtering out as much noise as possible. This involves both calculation of new variables as well as filtering out variables with low predictive value.

If feature engineering is performed parallel to building the final predictive model, there is a risk of inadvertently overfitting the features to the training data, meaning the final feature set will have good predictive value on the training set but bad generalization to unseen or external data. In order to avoid this problem, feature engineering will be done for a model predicting the traditional clinical response and remission criteria. This should be beneficial under the assumption, that variables with predictive value for the traditional criteria will also have predictive value for the cluster structure.

In this experiment, a set of models for predicting traditional response and remission criteria will be trained and evaluated. The predictive models will differ in the sets of predictive variables. The target variables will be response or remission by the traditional criteria in week 4. Based on the results, a set of variables to use for building a predictive model for the clusters will be selected in the discussion section. At the same time, this experiment expands upon previous data on the predictability of treatment response (see page 11).

Methods

Design Summary

A random forest classifier will be trained with different sets of predictor and target variables. For evaluation of classifier performance, 10-fold cross validation will be performed. This means, the classifier will be trained on 9/10th of the patients and will predict the last 1/10th of patients. This process will be repeated 10 times with distinct sets of patients used for prediction, so the entire dataset will have been predicted exactly once. Performance of the classifier models will be evaluated against random guessing as well the early improvement criterium.

Decision Tree Classifier

For more detailed information on decision trees, see [51], the source for the following explanation. A decision tree classifier is a classification technique used to predict, to which one of several distinct classes an observation belongs. This is achieved by chaining multiple (usually binary) decisions with each decision separating the chain into multiple (two in the binary case) branches. This leads to a structure of chained decision branches, thus the name decision tree. Each individual decision can be referred to as a node. The training algorithm uses a set of training data with known target classes to create the decision tree. Since the target classes are known, this is considered supervised machine learning. The algorithm begins by picking a variable and a value that achieves the best separation of the observations into the corresponding classes. The quality of separation can be measured in a multitude of ways, common ones include Gini-Impurity or Variance Reduction. Each resulting branch is then again separated in this manner, until a stop condition is met. Examples of possible stop conditions include, that only one class remains in the branch or that either the current branch or branches resulting of a possible split have too few elements. The last nodes of the decision tree are also referred to as leaf nodes. For an example visualisation of a decision tree, see Figure 24.

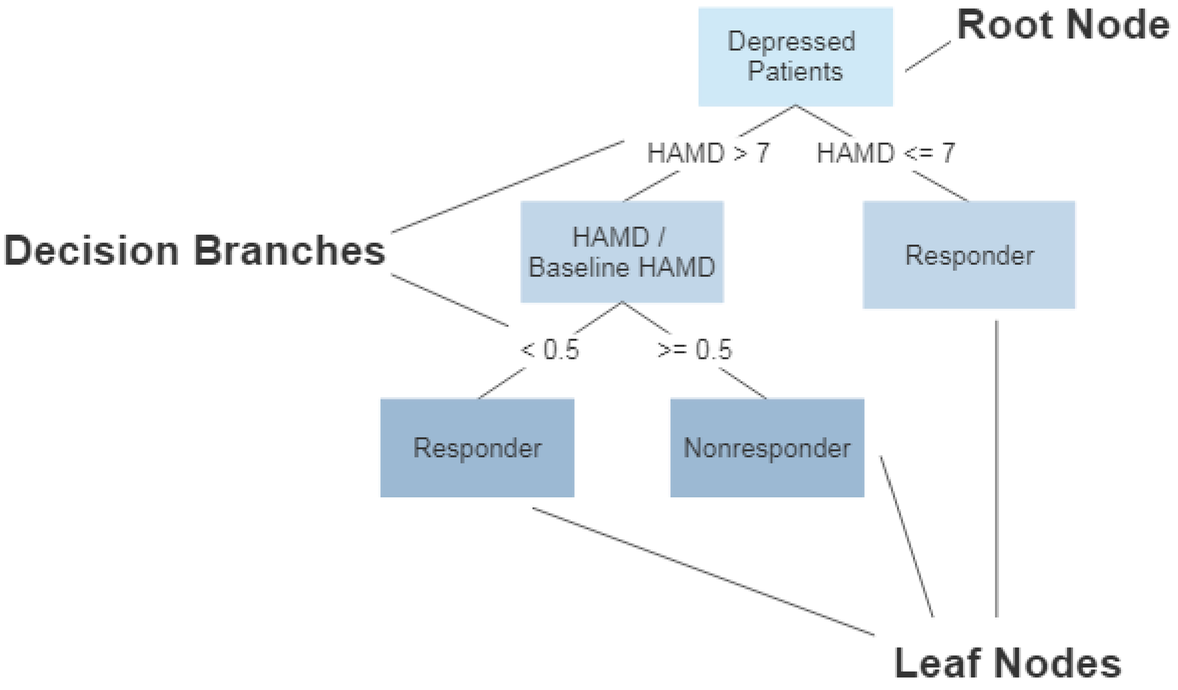


Figure 24: Example of a decision tree for the traditional response criterium based on the HAMD score. Parts of the decision tree are labelled.

Random Forest Classifier

For more detailed information on random forest classifiers, see [52], the source for the following explanation. A random forest classifier is an extension of the decision tree classifier described above. Instead of building a single decision tree, which often leads to poor generalization due to making decisions too specific to the training data by capturing noise (so-called overfitting), a random forest classifier builds multiple decision trees. Depending on the algorithm settings, each individual tree is built using only part of the training observations, part of the training features and/or is “pruned” to only reach a certain depth. These steps are taken, so multiple trees in the classifier aren’t the same. After multiple individual trees are trained, they can be used to predict an unseen example via majority voting of the results from the individual trees. A degree of certainty in the prediction can also be calculated by quantifying, how “unanimous” the best classification was between the different trees in the forest.

The importance of individual variables for a random forest classifier can be estimated by calculating the average decrease of the Gini coefficient over all trees or by calculating the effect on the accuracy if the variable is randomly shuffled between observations. Calculation of these importance scores is a key part of selecting features with large predictive value and eliminating features that constitute noise. The Gini-coefficient based importance score will be calculated for all sets of variables.

Cross-Validation

In order to calculate the prediction accuracy of the classifiers, a cross-validation procedure, specifically 10-fold stratified cross-validation, is used. The dataset is split into 10 parts with patients being stratified based on the target variable, meaning there will be approximately equal numbers of patients with or without the target variable across dataset parts. The classifier model is then trained 10 separate times, withholding a separate part of the dataset each repetition. The unseen dataset part for each model is then given to that model for prediction. By proceeding in this manner, there is a prediction for each observation by a model that has not seen this observation during training. These predictions can then be compared to the known target variables for each observation, so accuracy and other estimates of test performance can be calculated.

Transformation of variables

Since random forest classifiers are based on recursive partitioning of the dataset, the resulting classification does not change based on basic mathematical transformations on the input. Therefore, a “normalization” (e.g. transforming all variables to zero mean, unit variance) is not required and is thusly not performed here [53]. The algorithm also natively can deal with both categorical as well as continuous data, so no additional pre-processing is necessary regarding the mixed properties of the input data. For categorical data, it is noteworthy though, that importance scores can be overestimated for variables with many categories compared to variables with few categories. This should be kept in mind for later variable selection.

For the weekly psychometric values (HAMD and IDS) some basic variable transformations are performed to optimally extract a possible predictive value. First, relative sum scores values to baseline are calculated. Second, absolute sum score differences to all prior timepoints are calculated. Third, principle component analysis is performed at baseline for the items of each test separately. Components are kept so the total explained variance is above 50%. The individual values for each patient for these principal components are calculated until week 2 (itemized psychometrics aren't available in the dataset for weeks 3 and after). The variables resulting from these transformations will be referred to as feature engineering variables for the purpose of this thesis.

Prediction variable sets

To evaluate the predictive values of the different available variables, multiple predictive sets are chosen for comparison. Each variable set will be calculated in 4 different time versions for every week (baseline until week 3). It will encompass all available data until that timepoint. One variable set includes all information based on the HAMD (sum score, items, feature engineering variables) until the corresponding week. A second set will include both the self and clinician rated IDS information. The third set will combine both the HAMD and IDS set. And a last set will combine HAMD and IDS information and add information from the M.I.N.I., SCID-II, demographics as well as medical and psychiatric history. This set is referred to as "All Variables" set.

Hyperparameters

Algorithmic settings are commonly called Hyperparameters. Random forest classifiers have multiple hyperparameters, like the number of trees in the forest, the type of information gain criterion (gini coefficient or entropy gain), the maximum decision tree depth or the maximum features considered per tree. These hyperparameters are commonly set using so called hyperparameter grid search, meaning that a space of parameter settings is searched for an optimal solution by calculating the cross-validation score for every possible combination of these parameters. In order to prevent contamination of the global cross-validation scores used for evaluation, this grid search is commonly done separately for each cross-validation fold. In order to compare the accuracies of different hyperparameter settings, cross validation will be employed again, though only on the training data of the respective cross validation fold. This is called nested cross validation hyperparameter tuning. For the nested cross validation, 3 folds are used.

In order to make this procedure computationally viable, the hyperparameter search space needs to be narrowly defined. Every possible parameter setting needs to be evaluated in combination with all other settings on all other parameters – for every cross-validation fold. The number of individual models that need to be fit quickly grows too large to calculate. Keeping this in mind, the search space for hyperparameter tuning is defined by 3 parameters: the information gain criterion (gini coefficient vs. entropy gain), the maximum decision tree depth (50, 100, Unlimited) and the maximum features considered per tree (square root of total features, log2 of total features, 50% of total features). The number of trees in the forest is fixed to 2000, which is likely larger than required, though it is computationally cheaper to estimate this larger number once for every parameter, than to evaluate all parameter combinations against multiple settings for the number of trees. This leaves a total of 18 parameter combinations, evaluated on 10 cross validation folds for 4 variable sets over 4 timepoints for 2 target variables, with evaluation being based on 3 folds per parameter combination, meaning a total of 18*10*4*4*2*3 = 17280 models will have to be fit (not including the final model fit with the best set of hyperparameters). At 5 seconds per model, this will take approximately 24 hours, showing how quickly a larger search space would become calculation-cost prohibitive.

Feature importance calculation

Due to the cross-validation procedure described above, there will be multiple models per set of predictive variables. In order to allow interpretation of the feature importance scores (see the description of random forest classifiers above, page 59), it is beneficial to have a single importance score per variable set. Therefore, the importance scores will be averaged across all models resulting from the same variable set.

In order to compare importance scores across multiple variable sets, it is important to keep in mind that simple averaging might not result in the intended effect. Importance scores need to be compared over time, since it seems likely that variables very important for baseline prediction might differ from variables very important for prediction at week 3. Furthermore, it should be considered, that the number of sets a given variable is in differs. Feature importance scores (after averaging over the models per variable set) are therefore ordered by rank, separately for every variable set and timepoint. If a feature shows a high rank (high meaning most important here) in multiple variable sets, the predictive value is likely real. If it ranks high in some variable sets but low in others, this might be due to cross-correlation between variables in the sets. If it consistently ranks low across variable sets, it likely captures noise and should be excluded from the set of predictive variables for future analysis.

Results

Prediction performance metrics

After calculation of the predictive models, standard binary confusion matrices and their corresponding metrics can be calculated. The accuracies for the given variable sets over time are shown in Figure 25 (Response at week 4) and Figure 26 (Remission at week 4). An overview of the remaining prediction metrics as well as Chi²-Tests and effect sizes is given by Table 10 (Response at week 4) and Table 11 (Remission at week 4).

For response at week 4, maximum prediction accuracy at baseline is 52.2%, only slightly above the zero-information rate (approx. 50%) with a steep increase in prediction accuracy to between 66.0% and 67.8% across all variable sets at week 1. Until week 3, all variable sets show a further, roughly linear, increase of accuracy to between 75.8% to 79.7%. The 20% early improvement criterium shows a roughly constant prediction accuracy between 63.2% and 65.7%, the 30% early improvement criterium shows an increase in accuracy from 67.0% at week 1 to 72.7% at week 3. Both sensitivity and specificity are roughly equal to each other for all variable sets and timepoints. Chi²-Testing against random guessing shows highly significant p-values for weeks 1 through 3 for all variable sets with non-significant values across all variable sets at baseline. Effect-size of the prediction as estimated with Cramer´s V increases monotonously over time for all variable sets.

For remission at week 4, prediction at baseline is distributed between 69.2% and 70.4% around the zero-information rate (approx. 70%) for all variable sets, with the set including all variables being worse than the others at approx. 62.1%. From there, all variable sets show a quasi-linear increase in predictive accuracy until week 3, where the accuracies are distributed between 83.4% and 86.3%. Both 20% and 30% early improvement criterium show a predictive accuracy lower than the zero-information rate at all timepoints. For all variable sets and timepoints, specificity is at least moderately better than sensitivity. Chi²-Testing shows significant p-values over random guessing for all timepoints and variable sets with the exception of the All-Variables-Set at baseline, that shows a non-significant value. Effect-size of the prediction as estimated with Cramer´s V increases monotonously over time for all variable sets.

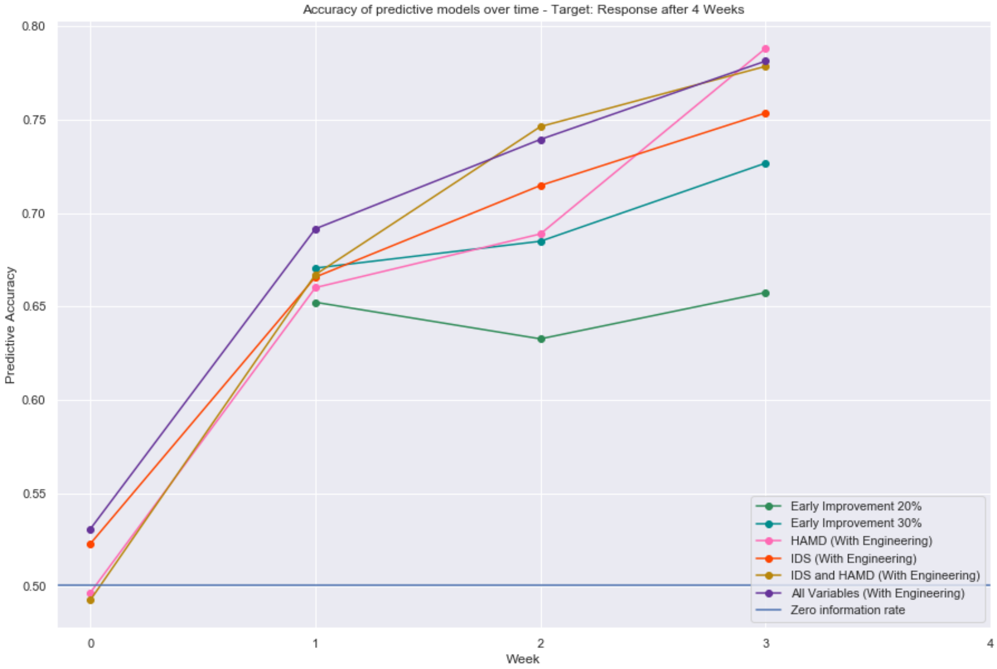


Figure 25: Cross-validation accuracy of random forest classifiers predicting the traditional 50% response criterium based on the given variable set at different timepoints. Zero information rate and early improvement criterium are shown for comparison.

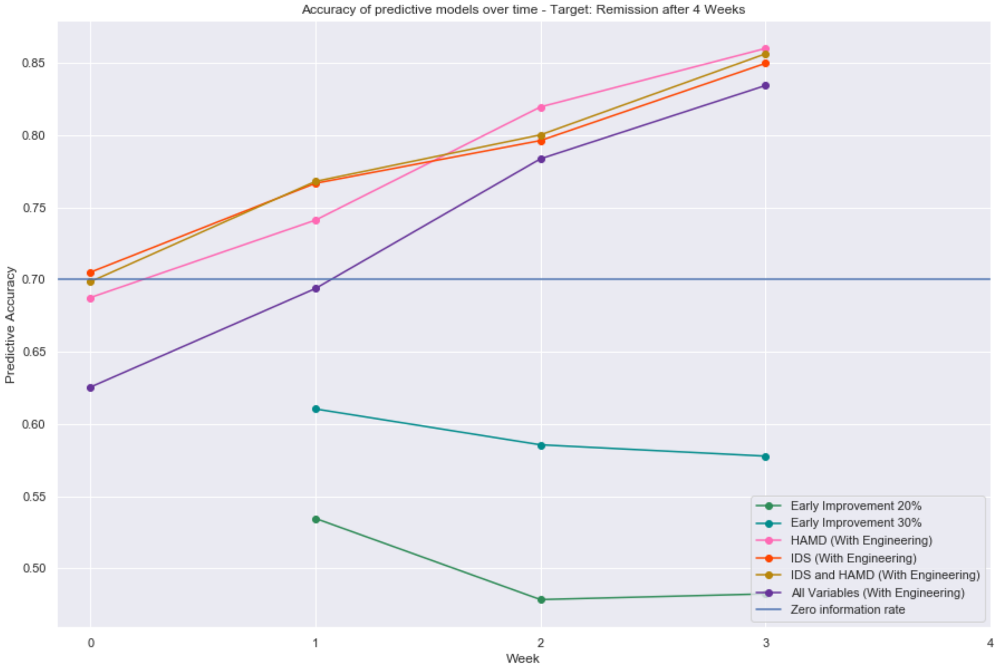


Figure 26: Cross-validation accuracy of random forest classifiers predicting the traditional remission criterium based on the given variable set at different timepoints. Zero information rate and early improvement criterium are shown for comparison.

Metric	Baseline	Week 1	Week 2	Week 3
All Variables as predictors for Response at week 4.				
Accuracy	52.2%	67.8%	74.0%	79.7%
Sensitivity	53.9%	68.7%	75.6%	80.6%
Specificity	49.1%	66.8%	72.1%	78.7%
Chi²-p Value	0.608	< 0.001	< 0.001	< 0.001
Cramer's V	0.024	0.349	0.473	0.588
HAMD Variables as predictors for Response at week 4.				
Accuracy	50.7%	66.0%	68.9%	79.3%
Sensitivity	50.8%	66.6%	69.6%	79.5%
Specificity	50.7%	65.5%	68.2%	79.2%
Chi²-p Value	0.745	< 0.001	< 0.001	< 0.001
Cramer's V	0.012	0.318	0.375	0.584
IDS Variables as predictors for Response at week 4.				
Accuracy	51.9%	66.1%	70.6%	75.8%
Sensitivity	52.0%	66.1%	71.5%	76.8%
Specificity	51.8%	66.0%	69.7%	74.7%
Chi²-p Value	0.331	< 0.001	< 0.001	< 0.001
Cramer's V	0.035	0.318	0.409	0.513
HAMD and IDS Variables as predictors for Response at week 4.				
Accuracy	50.1%	66.2%	73.9%	77.7%
Sensitivity	50.3%	66.4%	75.1%	78.5%
Specificity	49.9%	66.0%	72.7%	77.0%
Chi²-p Value	0.970	< 0.001	< 0.001	< 0.001
Cramer's V	0.001	0.321	0.475	0.552

Table 10: Binary classification metrics of random forest classifiers with the given variable sets as predictors and the traditional 50% response criterium at week 4 as target variable. Chi-2-Metric and Cramer´s V were calculated in comparison to random guessing.

Metric	Baseline	Week 1	Week 2	Week 3
All Variables as predictors for Remission at week 4.				
Accuracy	62.1%	71.4%	77.9%	83.4%
Sensitivity	36.0%	61.9%	70.7%	79.0%
Specificity	65.3%	74.7%	81.2%	85.5%
Chi^2-p Value	0.975	< 0.001	< 0.001	< 0.001
Cramer’s V	0.001	0.331	0.498	0.624
HAMD Variables as predictors for Remission at week 4.				
Accuracy	69.2%	73.3%	82.1%	86.3%
Sensitivity	46.0%	58.5%	76.4%	79.8%
Specificity	72.1%	76.9%	83.8%	88.7%
Chi^2-p Value	< 0.001	< 0.001	< 0.001	< 0.001
Cramer’s V	0.121	0.301	0.548	0.662
IDS Variables as predictors for Remission at week 4.				
Accuracy	70.4%	76.7%	80.0%	84.8%
Sensitivity	51.8%	66.0%	69.7%	78.5%
Specificity	72.6%	79.5%	83.6%	87.1%
Chi^2-p Value	< 0.001	< 0.001	< 0.001	< 0.001
Cramer’s V	0.162	0.400	0.504	0.627
HAMD and IDS Variables as predictors for Remission at week 4.				
Accuracy	70.2%	77.1%	80.4%	85.6%
Sensitivity	51.2%	67.5%	71.3%	80.3%
Specificity	72.7%	79.5%	83.4%	87.5%
Chi^2-p Value	< 0.001	< 0.001	< 0.001	< 0.001
Cramer’s V	0.160	0.408	0.511	0.646

Table 11: Binary classification metrics of random forest classifiers with the given variable sets as predictors and the traditional HAMD remission criterium at week 4 as target variable. Chi-2-Metric and Cramer´s V were calculated in comparison to random guessing.

Feature importance

For reasons of brevity, full feature importance ranks, values and graphs are not reported here. Instead, only feature importance graphs for the top 30 predictors for Response at week 4 in the All-Variable-Set are reported for baseline, week 1, week 2 and week 3 in Figure 27, Figure 28, Figure 29 and Figure 30, respectively. Additional graphs can be found in the appendix. Rank comparisons are reported for clinical history and demographic variables as to their predictive importance for Response at week 4 in Table 12, in order to facilitate discussion of possible inclusion in future models.

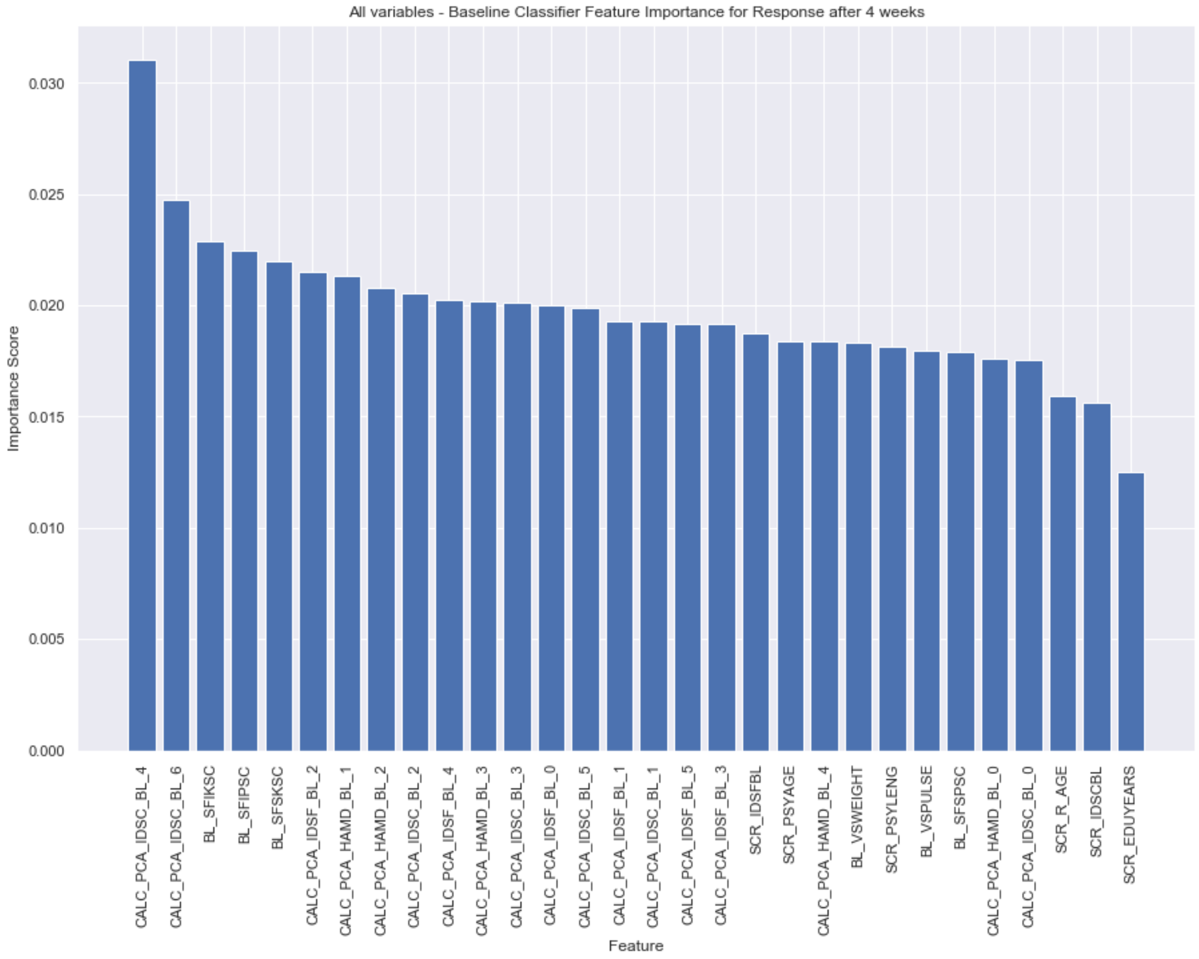


Figure 27: Feature importance for the prediction of response at week 4 in the All-Variables-Classifier at baseline. Variable abbreviations: SCR_*_(_rel): HAMD or IDSC (IDS clinician rated) or IDSF (IDS self-rated) score at stated). Added suffix "_rel" to indicate value relative to baseline. CALC_PCA_*_*_*: Principle component analysis of HAMD, IDSC or IDSF items at stated timepoint, nth component (0-indexed). CALC_DELTA_*_*_*: Score difference of HAMD, IDSC or IDSF between two stated timepoints. BL_SF**SC: SF-12 (I: Clinician rated, S: self-rated) (P: mental, k: physical) subscale score at baseline. SCR_PSYAGE: Age at first MDD episode. SCR_PSYLENG: Length of current MDD episode. SCR_R_AGE: Age at baseline. SCR_EDUYEARS: Years of formal education. BL_VSWEIGHT: Body weight at baseline. BL_VSPULSE: Heart rate at baseline.*

At baseline, where prediction accuracy of the All-Variables-Set classifier predicting traditional 50% response at week 4 was not significantly above random, the features with the highest importance scores were principal components of both HAMD and IDS Items, baseline SF-12 subscale scores as well as some clinical (Weight, Heartrate), history (Length of current episode, Age at first episode) and demographic variables (Years of education, Age).

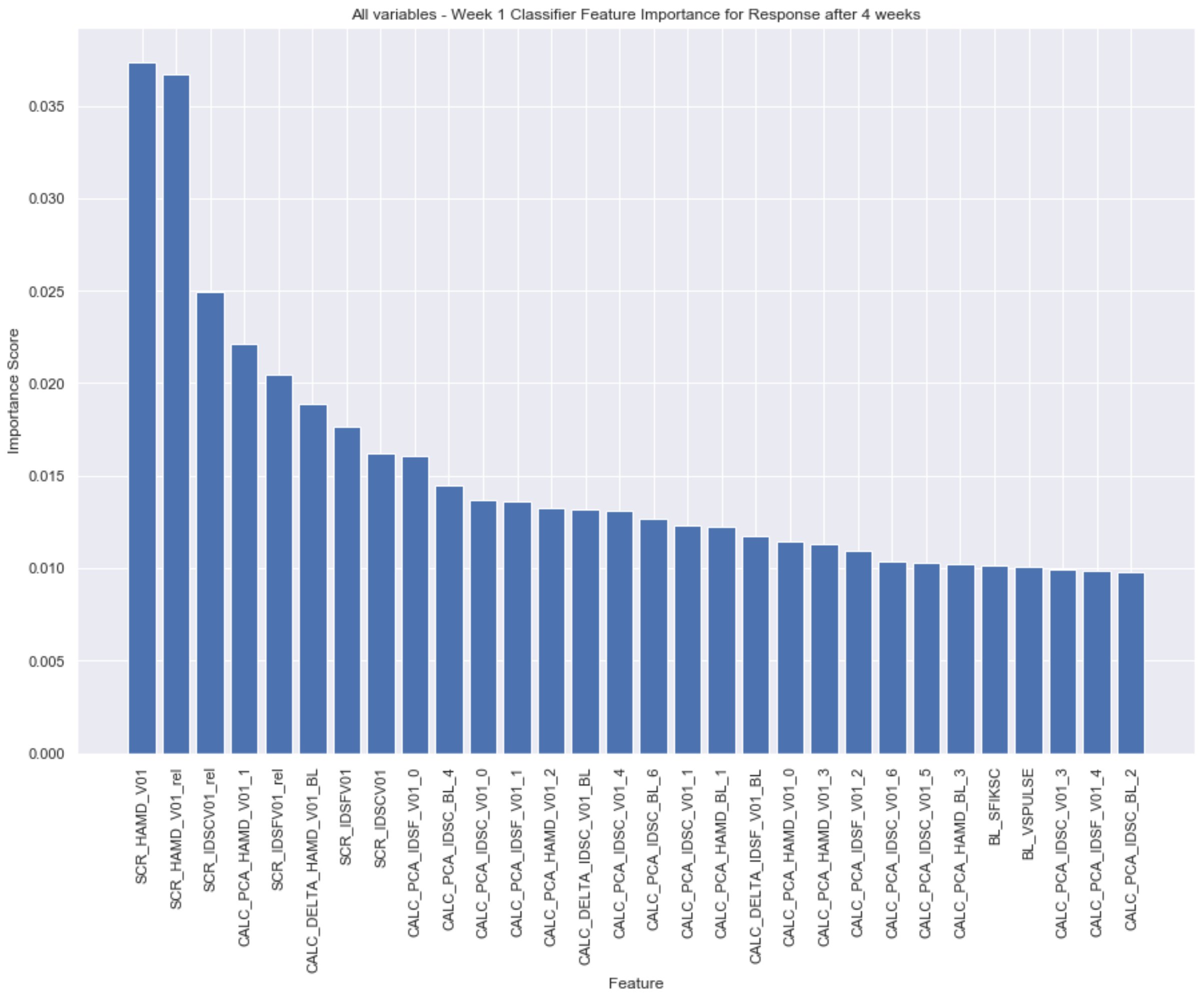


Figure 28: Feature importance for the prediction of response at week 4 in the All-Variables-Classifier at week 1. Variable abbreviations: SCR_*_(_rel): HAMD or IDSC (IDS clinician rated) or IDSF (IDS self-rated) score at stated). Added suffix "_rel" to indicate value relative to baseline. CALC_PCA_*_*_*: Principle component analysis of HAMD, IDSC or IDSF items at stated timepoint, nth component (0-indexed). CALC_DELTA_*_*_*: Score difference of HAMD, IDSC or IDSF between two stated timepoints. BL_SF**SC: SF-12 (I: Clinician rated, S: self-rated) (P: mental, k: physical) subscale score at baseline. BL_VSPULSE: Heart rate at baseline.*

The All-Variables-Set classifier predicting traditional 50% response at week 4 based on week 1 variables showed both the absolute and relative HAMD sum score at week 1 as having the highest feature importance. Other important variables included IDS sum scores, sum score differences between week 1 and baseline, principal components of psychometric scale items at baseline and week 1 and both baseline heart rate and SF-12 subscales.

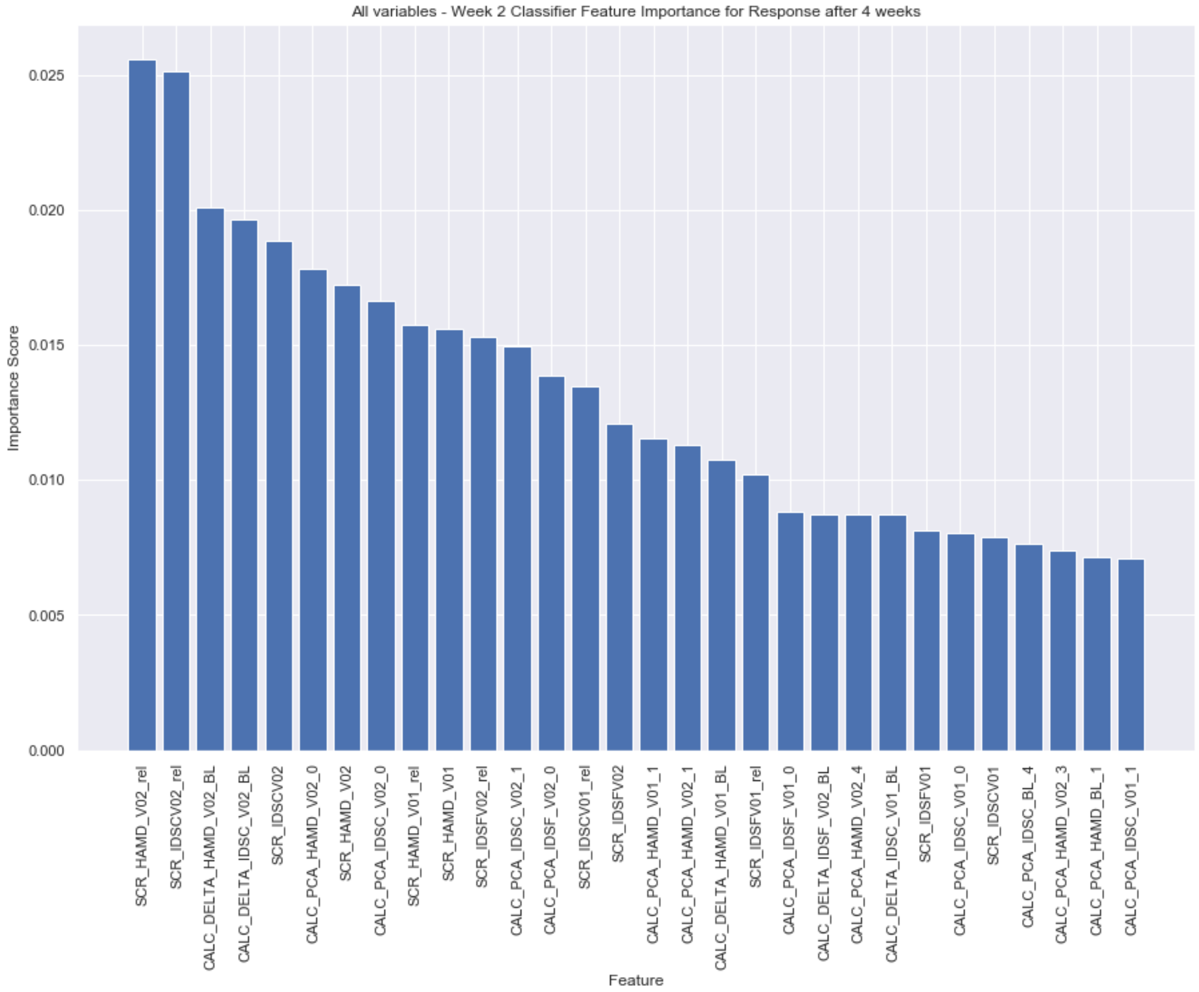


Figure 29: Feature importance for the prediction of response at week 4 in the All-Variables-Classifier at week 2. Variable abbreviations: SCR_*_(_rel): HAMD or IDSC (IDS clinician rated) or IDSF (IDS self-rated) score at stated). Added suffix "_rel" to indicate value relative to baseline. CALC_PCA_*_*_*: Principle component analysis of HAMD, IDSC or IDSF items at stated timepoint, nth component (0-indexed). CALC_DELTA_*_*_*: Score difference of HAMD, IDSC or IDSF between two stated timepoints.*

The most important features in the week 2 All-Variables-Set classifier predicting traditional 50% response at week 4 were relative HAMD and IDS sum scores at week 2, differences between these scores and baseline, sum scores at week 1 and differences to baseline and principal components of psychometric items between baseline and week 2.

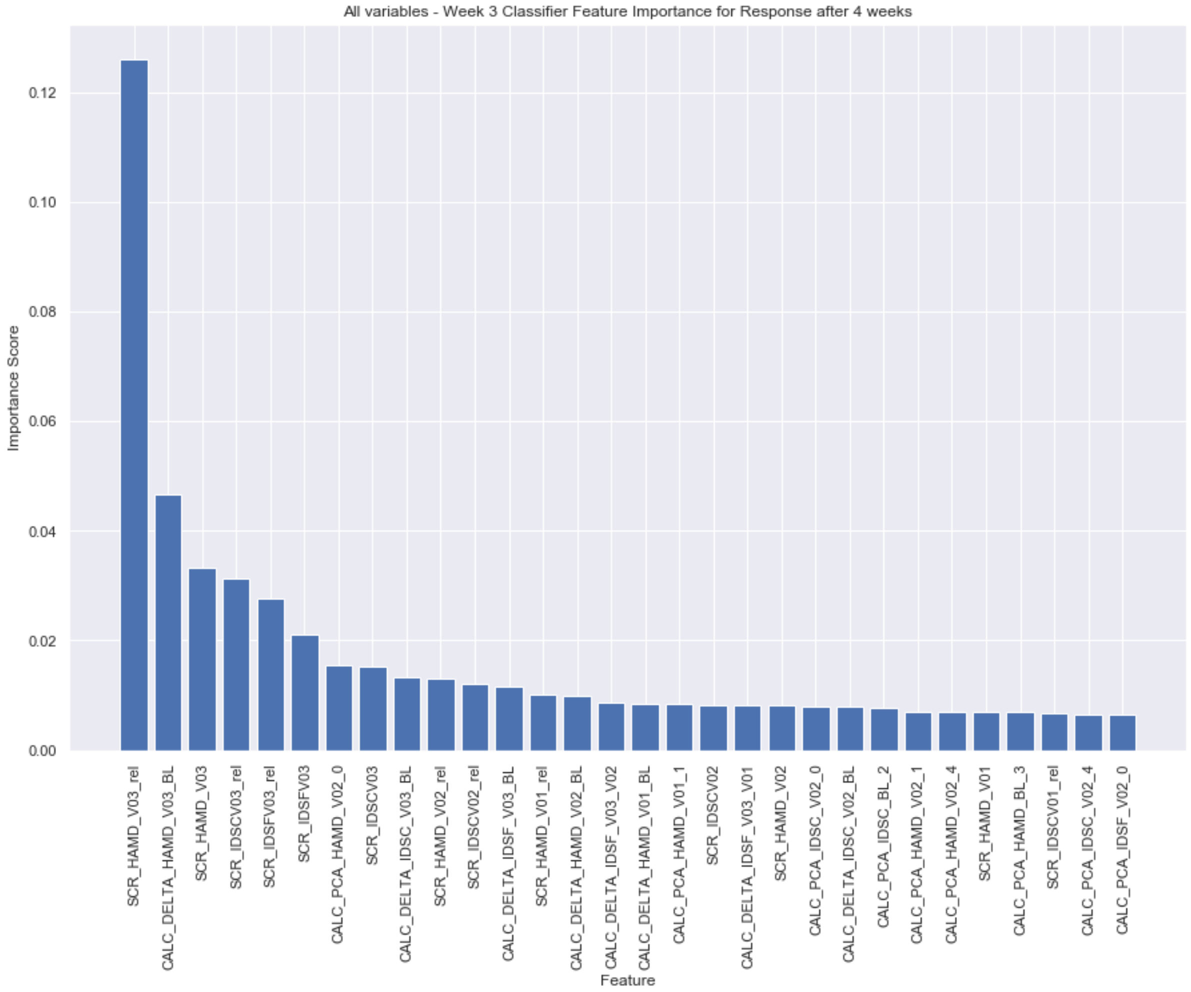


Figure 30: Feature importance for the prediction of response at week 4 in the All-Variables-Classifier at week 3. Variable abbreviations: SCR_*_(_rel): HAMD or IDSC (IDS clinician rated) or IDSF (IDS self-rated) score at stated). Added suffix "_rel" to indicate value relative to baseline. CALC_PCA_*_*_*: Principle component analysis of HAMD, IDSC or IDSF items at stated timepoint, nth component (0-indexed). CALC_DELTA_*_*_*: Score difference of HAMD, IDSC or IDSF between two stated timepoints.*

For the week 3 All-Variables-Set classifier predicting traditional 50% response at week 4, the relative HAMD sum score at week 3 shows a far bigger feature importance than all other features. Other important features are absolute and relative HAMD and IDS sum scores at weeks 1 through 3, their differences to baseline and principal components of psychometric scale items from baseline through week 3.

Variable	Baseline	Week 1	Week 2	Week 3
Age at first MDD episode	20 / 140 (14.3%)	53 / 216 (24.5%)	78 / 296 (26.4%)	79 / 312 (25.3%)
Length of current MDD episode (days)	23 / 140 (16.4%)	35 / 216 (16.2%)	82 / 296 (27.7%)	78 / 312 (25.0%)
Age at baseline	28 / 140 (20.0%)	58 / 216 (26.9%)	92 / 296 (31.1%)	75 / 312 (27.2%)
Years of education	30 / 140 (21.4%)	62 / 216 (28.7%)	90 / 296 (30.4%)	94 / 312 (30.1%)
Number of previous MDD episodes	32 / 140 (22.9%)	70 / 216 (32,4%)	111 / 296 (37.5%)	119 / 312 (38.1%)
Type of school education	50 / 140 (35.7%)	102 / 216 (47.2%)	147 / 296 (49.7%)	181 / 312 (58.0%)
Type of vocational education	56 / 140 (40.0%)	140 / 216 (64.8%)	175 / 296 (59.1%)	215 / 312 (69.2%)
Gender	93 / 140 (66.4%)	173 / 216 (80.1%)	252 / 296 (85.1%)	249 / 312 (79.8%)
Ethnicity	108 / 140 (77.1%)	180 / 216 (82.9%)	267 / 296 (90.2%)	281 / 312 (90.1%)

Table 12: Feature importance ranks of demographic and clinical history variables form the All-Variables-Set classifiers for the given timepoints predicting response at week 4. Ordered form Rank 1 = Most important. Percentage Ranks given in brackets.

Demographic and clinical history variables showed varying levels of feature importance for the All-Variables-Set classifier predicting traditional 50% response at week 4 with their relative importance ranking of the variables being roughly consistent across all timepoints. Age at first MDD episode, Length of current episode, Patient Age and Years of Education all consistently rank in the top third of feature importance. The number of previous MDD episodes shows a decline in importance from the 22.9th percentile at baseline to the 38.1st percentile at week 3. Gender, Ethnicity and type of education don´t rank in the top third at any timepoint.

Discussion

Conclusions and deductions from model prediction accuracy over time

The overview graphs of the predictive model accuracy over time (Figure 25 and Figure 26) are immediately suggestive of several key findings. Firstly, prediction at baseline seems to be at best moderately above the zero-information rate. Second, prediction gets more accurate closer to the target timepoint. And third, in the case of response prediction, the growth in prediction accuracy between baseline and week 1 strongly differs in magnitude from the growth after that point. These findings will be discussed in the following subsection.

While the baseline prediction of remission at week 4 beats random guessing with statistical significance in 3 out of 4 models, the prediction is at best a small amount above the zero-information rate. For prediction of response, none of the models are significantly better than random or show accuracy far above the zero-information rate. At first glance, this seems to be differing from previous results like the predictors of response discussed above (see page 11). But single predictors being associated with response at baseline don´t necessarily translate into prediction being possible, because response is most likely influenced by multiple factors and possibly interactions between these factors. If this relationship is either to complex to uncover for the chosen algorithm with the given amount of training data or the training data has a low signal to noise ratio prohibiting finding of this relation, prediction might not be possible even though the data contains variables already known to be predictive in some capacity.

One way to avoid this problem would be to increase the signal to noise ratio by feature selection, meaning only the most predictive features will be selected in order to build a model. This was done by Chekroud et al. (2016), where they achieved a prediction accuracy of approx. 60% on the external validation dataset by selecting only the top 25 predictors [31]. The use of an external validation set is key for this approach. Selecting the most informative features for the entire dataset and then trying to estimate the prediction accuracy via cross validation on the same dataset constitutes a so called “data leak”: Because it is already known from feature selection, that the selected features are valuable to predict the holdout part of the dataset, it is unsurprising that the holdout part gets predicted well by these features – the cross-validation accuracy will likely be overestimated and in order to get a “real” accuracy, an external validation is required. Chekroud et al. (2016) achieved a cross validation accuracy of 64.6% and an external validation accuracy of 59.6%, showing this effect [31].

For the purposes of this experiment, there was no external validation dataset available. Thus, feature selection on the dataset could not be employed in this manner in order to raise the signal to noise ratio and by doing so possibly raise the prediction accuracy above the zero-information rate. For Chekroud et al. (2016), the zero-information rate was 51.3% [31]. Even with employing feature selection in the manner described above, they achieved an external classification accuracy of only 59.6%. Compared to the zero-information rate of 50.0% for response prediction and a prediction accuracy of 52.2% for the best predictive model at baseline in this thesis, this is obviously better but doesn´t – in the authors opinion – justify the assumption of a difference not explained by the difference in methodology and the differing prediction question (Chekroud et al. (2016) predicting remission after 12 weeks vs. this experiment predicting response or remission after 4 weeks [31]). Overall, both in previous literature and the current experiment baseline prediction of treatment outcome based on clinical data alone seems to be possible at best moderately above the zero-information rate.

Unsurprisingly, the prediction accuracy increases with variables from timepoints closer to week 4 being used. This result serves as a sanity check of the prediction accuracy. What is interesting, is that in prediction of remission, this increase is approximately linear (at least when considering an average of all 4 variable sets) from baseline until week 3 and in prediction of response it´s linear from week 1 until week 3. This approximately linear increase can serve as an argumentative "baseline" about how much easier prediction gets over time. With prediction of response showing a much larger increase in accuracy between baseline and week 1 than suggested by the baseline linear relation from the rest of the data, closer attention is warranted.

The background section on early improvement (See page 10) details, that significant differences between responders and non-responders can be found within one or two weeks. This is indicative of an onset of antidepressant response within this timeframe. The difference in magnitude between the accuracy increase in comparison to a baseline linear relation could be interpreted as this onset of response. In this case, the data from the current experiment indicates an onset of antidepressant response within the first week, in line with results from Szegedi et al. (2003) [23].

This interpretation raises the question, why the onset is only seen in the predictive accuracy of response prediction and not seen in remission prediction. One possible explanation is in the timeframe of the remission prediction. The current experiment predicted remission after 4 weeks, which is earlier than remission would be expected by current antidepressant treatment strategy (See page 9), that expects response by week 4 to be predictive for later remission. But the shorter the distance to the predicted timepoint, the stronger the accuracy gain over time should self-evidently be. Therefore, it could be possible for this strong accuracy gain over time to mask the presumed accuracy gain from an onset of response. Since in the EMC trial, there were antidepressant treatment adaptations directly dependent on whether a patient showed response at week 4, predictions for remission in the further time course of this dataset would change the prediction question sufficiently to prohibit investigation of this possibility: An onset of antidepressant response to one medication would not necessarily show up in the prediction accuracy for an entire treatment strategy. This analysis should be done in future research on a different dataset, where patients were exposed to the same antidepressant treatment for longer, since the current results could indicate a way of supporting the early onset of antidepressant response further and quite specifically.

Comparison to the early improvement criterium

In comparison to both the 20% and 30% forms of the early improvement criterium, the machine learning models outperform. The 20% early improvement criterium shows a predictive accuracy of approx. 65% for the prediction of response at all timepoints, which is worse than the models based on all variable sets at week 1 and much worse at all other timepoints. The 30% early improvement criterium shows a predictive accuracy on par with the machine learning models in week 1 and an increase in accuracy over time, but this increase can´t match the accuracy increase of all other models that outperform the early improvement criterium in both week 2 and week 3 for the prediction of response. For the prediction of remission at week 4, both early improvement criteria are essentially worthless at all timepoints, since they are far below the zero-information rate.

In addition to the early improvement criterium being outperformed accuracy-wise, the machine learning models have several further advantages. As described in the methodology for random forests (See page 59), an estimate of prediction confidence based on the unanimity of predictions between trees can be made. While this was not evaluated in this experiment, it provides an additional source of information that could be considered for clinical decision making. Secondly, a trade-off between sensitivity and specificity can be made for the specific needs such models might be used for. The precision data in this experiment was automatically chosen to maximize accuracy, but unlike the cut-off for the early improvement criterium, this can be adjusted.

Feature selection for future model building

For beginning feature selection, the first key result is that models based on data from all variable sets perform on roughly the same level across all timepoints. Both IDS and HAMD seem like valid ways to measure symptom severity and neither approach seems to show a definitive benefit over the others. Since strong levels of collinearity can be expected between the different measures of symptom severity, only one set should be included. A strong argument for then choosing the HAMD score instead of the IDS score is present in the prediction question the features are selected for: The cluster structure was based on the HAMD score; therefore, it is justified to assume more contained information over the IDS. Since the clustering was specifically done on the relative HAMD scores, all relative HAMD scores should be included. This is mathematically justified, since the relative HAMD score is among the top predictors of the all variables model from weeks 1 through 3. In addition, the absolute level of symptom severity from all timepoints should be included. Absolute sum scale scores from both the latest and previous timepoints are present in the top 30 predictors of response for all timepoints of the All-Variable set model. This provides a mathematical justification for inclusion in addition to the theoretical argument of predictive information being contained in the time course of treatment response.

Regarding other features derived from the psychometric scales, it is obvious from the feature importance data of all timepoints, that principal components are more predictive than individual psychometric items. It is especially noteworthy, that principal component expressions from previous timepoints are still among the top predictive features in models from later timepoints. Further, it is noteworthy that the principal components´ predictive value is not present in rank order (meaning the first principle component is not the most important feature etc.). It therefore is appropriate to include all the HAMD principal components and exclude all single items variables.

Non-psychometric derived features are only included in the All-Variables set. The key features here are the characteristics from clinical variables and demographics. The theoretical importance of these variables was explored in detail in both the background section and the discussion section to experiment 1 (See page 11 and 52 respectively). The variables falling within the top third of predictive variables for at least one timepoint will be included, meaning inclusion of patient age, age of onset, length of current MDD episode, number of previous MDD episode and years of education. All other demographic variables will be excluded. Noteworthy here is the exclusion of ethnicity, which was an important feature for Chekroud et al. (2016) [31]. The EMC dataset features an approximately homogenously European population (See page 22). This might lead to ethnicity being less predictive in the current dataset than suggested by previous research, simply because the number of Non-European patients is too small to uncover a possible effect.

Additional Non-demographic features that show up within the top predictive features are SF-12 subscale scores at baseline and heartrate at baseline. Because of their relative importance to prediction of response even in later timepoints (e.g. top 30 predictors for the week 1 model), these features are included in addition.

To summarize, the following features will be included as predictive variables for the cluster prediction model: Absolute and relative HAMD sum-scores, HAMD item principal components, patient age, age of onset, length of current MDD episode, number of previous MDD episode and years of education, SF-12 subscale scores at baseline, heartrate at baseline.

Experiment 3 – Prediction of response type clusters

Summary of Experiment 3

Random forest classifiers were used to predict later assignment of patients to the clusters from Experiment 1 with variables available at Baseline and weeks 1 through 3 from the variable set selected in Experiment 2. Methodology for training and prediction assessment was the same as for Experiment 2.

Prediction accuracy increased over time, with baseline prediction not being better than random or the zero-information rates for any cluster. At week 1, predictions for the Early Response (0) cluster were better than the zero-information rate and random. At week 2, predictions for all clusters were better than random and predictions for all clusters except the Delayed Improvement (2) cluster were better than the zero-information rate. At week 3, all clusters were predicted significantly above random and better than the zero-information rate with classification accuracies between 86.2% and 95.5%. Specificity was high compared to sensitivity for all clusters at all timepoints. The high classification accuracy together with the high specificity implied possible clinical utility in early identification of a patient subpopulation for which early medication change might be beneficial. Limitations of this predictability and areas for future investigation of algorithmic performance were discussed.

Aim

After investigating prediction of traditional response markers for the purpose of feature selection, the extension of the predictive methodology from experiment 2 to the cluster structure from experiment 1 is the next step for deriving early clinical decisions based on the early treatment response patterns. Only if predicting the cluster assignment of a given patient early is possible with decent accuracy, early clinical decision making based on these predictions is a viable route for further investigation. If no such prediction is viable, only the time-course classification after 4 weeks is available as a new source of information for clinical insight (see page 16).

Method

Design Summary

A set of random forest classifiers with the selected set of predictive variables from the experiment 2 discussion will be trained with the different clusters from experiment 1 as binary target variables (patient belonging to cluster X or not). The predictive variable set will exist in 4 variants for the different prediction timepoints (baseline through week 3). Nested hyperparameter tuning will be employed for training these models. Accuracy, other test performance metrics and Chi^2-Tests against random will be evaluated based on 10-fold cross validation. Prediction test metrics will be evaluated in relation to the different timepoints. Feature importance will be calculated for all predictions.

Algorithmic methodology

The methodology for the random forest prediction in this experiment is identical for the prediction in experiment 2. For information on the underlying algorithm (see page 59), the cross-validation procedure (see page 60), the hyperparameter tuning and its search space (see page 61) and feature importance calculation (see page 62), please refer to the methodology section of experiment 2.

Predictive and target variables

The predictive variable sets and target variables differ from experiment 2. In the discussion from experiment 2, a set of predictive variables was selected (see page 75). These features are used as predictive variables here. The predictive variable set is given in 4 versions for the different possible timepoints of prediction (baseline through week 3). Each variable set includes all variables from its own and previous timepoints. The set of predictive variables includes absolute and relative HAMD sum scores, principal component expressions from PCA of HAMD items at baseline (from baseline through week 2, since no HAMD items were available for week 3), baseline SF-12 subscale scores, baseline heart frequency, the patients age, age at first depressive episode, length of current episode and years of education.

The target variables are Boolean variables indicating a patient's assignment to a given cluster. This can also be referred to as one-hot-encoding variables from the cluster variable. This means, each cluster is predicted one at a time. This methodology is opted for over a prediction of which cluster a patient will most likely be assigned to (predicting all 5 clusters at once) because of the underlying clinical implications. If there is a clinical hypothesis based on a patient's cluster prediction, it is likely uninteresting which cluster a patient belongs to. Instead, the key prediction is whether a patient belongs to the single cluster in question, so it can accurately be assessed, whether the hypothesis´ suggested intervention is indicated for the given patient.

Results

Prediction performance metrics

Timepoint	Early Response (0)	Early Improvement with Response (1)	Delayed Improvement (2)	Early Improvement without Response (3)	Non-Improvement (4)
Zero-Information Rate	80.1 %	69.3 %	83.3 %	79.7 %	87.6 %
Baseline	**80.3 %**	68.7 %	82.6 %	79.1 %	87.1 % **
Week 1	**85.3 % *** **	66.4 %	82.3 % *	79.1 %	87.5 % ***
Week 2	**91.5 % *** **	**76.2 % *** **	82.1 % **	**80.0 % *** **	**89.9 % *** **
Week 3	**95.5 % *** **	**87.9 % *** **	**87.9 % *** **	**86.2 % *** **	**94.6 % *** **

*Table 13 Accuracy of predictive model for the given cluster and timepoint. Stars indicate p-test resulting from Chi²-Testing vs. random *: p < 0.05, **: p < 0.01, *** p < 0.001. Zero-information rate given for additional comparison; bold text indicates values higher than the zero-information rate for the given cluster.*

An overview of prediction accuracies per cluster and timepoint is given in Table 13. For all clusters, accuracy at baseline approximately equals the zero-information rate. At baseline, only the Non-Improvement cluster (4) shows a significant p value in Chi²-Testing performance versus random. Accuracy at week 1 is still below the zero-information rate for all clusters with predictions for the Early Response (0), the Delayed Improvement (2) and the Non-Improvement (4) clusters now performing significantly better than random. Only the predictions for the Early-Response (0) cluster are above the zero-information rate. In weeks 2 and 3, predictions for all clusters perform significantly better than random. All predictions in this timeframe are also better than the zero-information rate except for the Delayed Improvement cluster (2) at week 2. Additional test performance metrics are shown in Table 14.

Metric	Baseline	Week 1	Week 2	Week 3
Early Response (0)				
Accuracy	80.3%	85.2%	91.5%	95.5%
Sensitivity	22.2%	64.2%	82.5%	91.8%
Specificity	81.0%	88.9%	93.2%	96.2%
Chi^2-p Value	0.857	< 0.001	< 0.001	< 0.001
Cramer’s V	0.007	0.479	0.705	0.844
Early Improvement with Response (1)				
Accuracy	68.7%	66.4%	76.2%	87.9%
Sensitivity	36.8%	29.4%	62.4%	83.2%
Specificity	69.5%	69.3%	81.4%	89.8%
Chi^2-p Value	0.732	0.968	< 0.001	< 0.001
Cramer’s V	0.012	0.001	0.421	0.707
Delayed Improvement (2)				
Accuracy	82.6%	82.3%	82.1%	87.9%
Sensitivity	0.0%	38.1%	38.5%	68.0%
Specificity	82.9%	83.6%	83.8%	91.1%
Chi^2-p Value	0.987	0.021	0.007	< 0.001
Cramer’s V	< 0.001	0.087	0.102	0.537
Early Improvement without Response (3)				
Accuracy	79.1%	79.1%	80.0%	86.2%
Sensitivity	28.6%	28.6%	54.0%	68.8%
Specificity	79.7%	79.7%	81.4%	90.0%
Chi^2-p Value	0.947	0.947	< 0.001	< 0.001
Cramer’s V	0.002	0.002	0.188	0.552
Non-Improvement (4)				
Accuracy	87.1%	87.5%	89.9%	94.6%
Sensitivity	50.0%	54.1%	65.2%	87.3%
Specificity	87.5%	89.4%	92.5%	95.4%
Chi^2-p Value	0.009	< 0.001	< 0.001	< 0.001
Cramer’s V	0.099	0.279	0.494	0.736

Table 14: Binary classification metrics of random forest classifiers with final selection variable set as predictors and the given cluster assignment as target variable. Chi-2-Metric and Cramer´s V were calculated in comparison to random guessing.

Feature importance

Feature importance values were calculated for all timepoints and all target clusters individually. The importance scores were averaged over all 5 clusters for each timepoint. Results are given in Table 15. Individual importance scores for the predictive models can be found in the appendix.

Variable	Baseline	Week 1	Week 2	Week 3
Rel. Week 3 HAMD	Not included	Not included	Not included	0.234
Rel. Week 2 HAMD	Not included	Not included	0.204	0.150
Rel. Week 1 HAMD	Not included	0.164	0.128	0.132
Abs. Week 3 HAMD	Not included	Not included	Not included	0.072
Abs. Week 2 HAMD	Not included	Not included	0.052	0.026
Abs. Week 1 HAMD	Not included	0.065	0.034	0.024
Abs. Baseline HAMD	0.055	0.036	0.022	0.016
HAMD PCA V02-0	Not included	Not included	0.034	0.022
HAMD PCA V02-1	Not included	Not included	0.031	0.019
HAMD PCA V02-2	Not included	Not included	0.025	0.014
HAMD PCA V02-3	Not included	Not included	0.027	0.016
HAMD PCA V02-4	Not included	Not included	0.025	0.015
HAMD PCA V01-0	Not included	0.042	0.022	0.014
HAMD PCA V01-1	Not included	0.055	0.029	0.022
HAMD PCA V01-2	Not included	0.048	0.024	0.015
HAMD PCA V01-3	Not included	0.039	0.021	0.014
HAMD PCA V01-4	Not included	0.040	0.023	0.013
HAMD PCA BL-0	0.080	0.043	0.025	0.015
HAMD PCA BL-1	0.073	0.041	0.024	0.014
HAMD PCA BL-2	0.073	0.038	0.022	0.014
HAMD PCA BL-3	0.068	0.037	0.023	0.014
HAMD PCA BL-4	0.065	0.036	0.022	0.012
SF-12-SR SSC	0.076	0.039	0.022	0.013
SF-12-SR PSC	0.071	0.034	0.020	0.013
SF-12-CR SSC	0.069	0.038	0.021	0.014
SF-12-CR PSC	0.070	0.037	0.022	0.014
Prev. Episodes	0.036	0.021	0.012	0.008
Age at Onset	0.061	0.033	0.019	0.011
Current Age	0.055	0.031	0.019	0.010
Length of Episode	0.058	0.031	0.018	0.011
Baseline Heartrate	0.046	0.026	0.015	0.008
Years of education	0.044	0.026	0.016	0.008

Table 15: Feature importance scores averaged over predictive models for all clusters at the given timepoints. Variable Abbreviations: Rel. Week X HAMD: Relative HAMD Score to baseline. Abs. Week X HAMD: Absolute HAMD score. HAMD PCA X-Y: Principle component analysis of HAMD items at timepoint X, component Y (zero indexed). SF-12 X Y: SF-12 score self-rated (SR) or clinician rated (CR) somatic subscale (SSC) or psychiatric subscale (PSC). Prev. episodes: Number of previous depressive episodes.

For prediction from baseline, the most important features were HAMD item principle components, followed by SF-12 sum-scores and lastly absolute HAMD score as well as demographic and clinical history information.

At week 1, the relative HAMD value from week 1 was the most important predictor by a large margin to the absolute HAMD score at week 1, the second most important predictor. After that, a group of variables including the HAMD item principle components for both timepoints, the absolute HAMD score from baseline and the SF-12 scores had roughly comparable average feature importance scores. The clinical history and demographic variables showed the lowest feature importance.

A similar pattern was found for weeks 2 and 3: All relative HAMD scores where the most important predictors by a large margin, with the latest absolute HAMD score being the next most important predictor. Next, the HAMD item principle components for all timepoints, the absolute HAMD score for all earlier timepoints and the SF-12 scores had roughly comparable average feature importance scores. Variables from clinical history and demographic data showed the lowest feature importance scores.

Discussion

Implications from the performance metrics

Prediction accuracy increases over time for all clusters. For week 2, predictions for all clusters are significantly better than random and at week 3, predictions for all clusters are both significantly better than random and above the zero-information rate. This distinction is important, as the target variables are highly imbalanced with zero-information rates between 69.4% and 87.6%. With labels that imbalanced, it is common to find results significantly above random, that have lower accuracy than the zero-information rate. This is easily explained when assuming a maximally imbalanced target variable with only 1 element that belongs to class A while all other elements belong to class B. The zero-information rate for this target is – depending on the cluster size – close to 100%. If all elements are assumed to be class B, this zero-information rate accuracy will be achieved. If the classes are randomly assigned based on their proportions, the likelihood of the single class A label to be correctly assigned is low. If class imbalance involves more than a single element but remains heavily skewed, this can lead to the phenomenon of predictions performing worse than the zero-information rate, while being better than random. Therefore, the zero-information rate is the better evaluator in case of the imbalanced class labels that were used.

The time course for individual clusters can be explained by their degree of separation from other clusters at the given time point. This can be approximated by the degree of separation of the cluster mean values as seen in Figure 22. Because the clustering algorithm from Experiment 1 takes all timepoints into account as individual variables, the degree of separation is cumulative over time. These interpretations are in line with the results from the current experiment. At baseline, no degree of separation exists between the cluster means and no results above random or above the zero-information rate. At week 1, only the Early Response (0) cluster has a large degree of separation from all other clusters. The Early Improvement clusters (1 and 3) have close to identical cluster means, as do the Non-Improvement (4) and Delayed Improvement (2) cluster. This implies a low degree of cluster separation at that point and subsequently, only the Early Response cluster can be predicted above random and zero-information rate at that timepoint. At week 2, some difference in cluster means exists between the two mentioned cluster pairs, leading to prediction results above random for all clusters and above the zero-information rate for all clusters except Delayed Improvement (2) cluster. At week 3, all cluster means show some degree of separation. The Early Improvement without response (3) and the Delayed Improvement (2) cluster do show a similar cluster mean for this timepoint, but these showed a difference in cluster means before then. All clusters therefore accumulated a sufficiently high degree of separation and as result, all clusters can be differentiated from each other with high accuracies above the zero-information rate.

Looking at the other predictive metrics, it is noteworthy that the increase in accuracy over time is mainly driven by an increase in sensitivity. Specificity is consistently high for all clusters at all timepoints. This is an expected result for an imbalanced target variable, but it has important clinical implications. Since a test with high specificity has few false positives, positive predictive value is generally high as soon as moderate sensitivity is reached (depending on the zero-information rate). This makes a test ideal for identifying a population, for which some intervention is indicated, when the associated risk of not correctly identifying this population is low. This is the case for early clinical intervention in patients with depression. When a patient can be identified as belonging to a cluster for which a hypothetical intervention (e.g. early change of medication) is indicated, this patient would be (on hypothetical average) less likely to receive medication for too long. If this patient is not correctly identified early, he receives treatment as usual with a change of medication after 4 weeks, which has no additional cost (when compared to all patients receiving treatment as usual). Since specificity of the cluster predictions from this experiment is high, these could – provided interventions can be shown to be effective in one or multiple of the clusters – be used as a test for identifying candidates for this intervention.

Limitations of predictability

The key result of this experiment was showing predictability of cluster assignment with high accuracies along all clusters. With predictive accuracies between 76.2% and 91.5% at week 2 and consistently good specificity, it seems viable to base clinical interventions on the predicted cluster assignment, though some limitations must be considered.

First, while the terminology of this thesis consistently used the term "prediction", the current experiment did not in fact predict cluster assignment for any timepoints after baseline. This is because the cluster assignment is directly based on the relative HAMD scores, which are given to the classification algorithm as variables. It is easy to imagine the classification algorithm achieving (near) perfect accuracy with all 4 relative HAMD values (classification based on full information). When only some of the relative HAMD scores are available, the algorithm similarly classifies based on partial information. Additionally, to this classification based on partial information, the algorithm utilizes information form the other available variables to improve its classification result. This can be conceptualized as the algorithm predicting the likely relative HAMD scores for the remaining timepoints. While mathematically inaccurate, this conceptualization is useful for distinguishing the partial classification task from the actual prediction task as two different information sources for the purposes of this discussion. The feature importance scores suggest, the partial classification variables (relative HAMD sum scores) to be much more important, than the other predictive variables. This difference in importance increases over time, which is consistent with the ratio of available to predicted information increasing over time. If distinction between these two information sources is required for future research, a possible way to test this would be to test the prediction vs. classification based on (partial) random walks.

Second, the accuracy in this experiment has a possibility of being overestimated due to data leakage. For performing variable selection on the results from Experiment 2, the entire dataset was used. Therefore, it is already known that the selected variables have some predictive significance for the entire dataset. By now evaluating a prediction based on these variables via cross-validation might introduce some overestimation based on dataset noise captured in the variable selection. Since this noise would not generalize to external data, the accuracy estimated by cross-validation might be too high. This overestimation was considered acceptable due to several factors. First, since variable selection and final prediction were done on two related, but different target variables, the amount of captured noise was considered to likely be low. Second, the process of variable selection including theoretical justification of the variable choices instead of purely accuracy-derived reasoning further decreases the risk of capturing noise. And third, considering the two aforementioned reasons, any remaining overestimation of accuracy was considered less important than the possible decrease in result interpretability due to lower dataset size. This refers to the possibility of avoiding this data leak by performing Experiment 2 and variable selection on only part of the dataset, while estimating accuracy in Experiment 3 on the remaining dataset (hold-out validation). Alternatively, a (dual) nested cross-validation approach could be used, though that was cost-prohibitive due to multiplicative computational cost increase for the already (single) nested cross-validation during hyperparameter tuning (See page 61).

A third limitation stems from the EMC dataset properties. Some patients without early improvement were randomized into the early medication change arm after 2 weeks, so not all patients received the same medication for the entire prediction period. And since the early medication change was randomized, it is inherently unpredictable from the datapoints before. This might somewhat prevent accurate prediction, especially for the Non-Improvement (4) and Delayed Improvement (2) clusters since these would be the clusters expected to have most of the randomized patients. This difficulty in predicting accurately would likely lead to an underestimation of prediction accuracy, which – considering the good prediction accuracy shown for later timepoints – is inconsequential for result interpretability. The good prediction accuracy also retroactively supports the decision from Experiment 1, that randomized patients were not excluded from analysis. It was assumed there, that a higher number of patients was more beneficial to overall prediction accuracy than the difficulties arising from the randomization.

Combined Discussion

Deriving hypotheses

The goal of this thesis was to create data-driven hypotheses to improve clinical decision making in the first 4 weeks of antidepressant treatment response (see page 16). In order to generate these hypotheses, it was attempted to identify clusters of treatment response over time (Experiment 1) and to predict and/or partially predict these patterns of response (Experiments 2 and 3).

Utility vs. mechanistic claim

Keeping this overarching goal in mind, it is important to discuss what precisely was investigated by deriving the cluster structure in Experiment 1. The aim of that experiment was to stratify patients by their respective early treatment response patterns in order to identify possible information for clinical decision making. The cluster structure proposed here should be understood as a mathematically optimized way to capture the room of possible response patterns, with the mathematical validity of this capture being shown by the goodness of fit measures discussed in detail above (see page 46). The good generalization to the validation dataset should also not be taken in the sense of “both datasets having the same clusters” but as “the cluster structure from the training set is a valid way of describing the patterns found in the validation dataset”. If training and validation dataset were reversed, the cluster structure would likely look different and result in another valid description.

Testing whether the early treatment response patterns are a valid description of the external validation data set is nevertheless a key step before investigating the cluster structure further, especially considering properties of the EMC dataset. Not all patients that were grouped into the patterns of early treatment response received the same medication for all 4 weeks. Some of the patients were randomized into the early medication change group after 2 weeks. These patients could be excluded from analysis, but since the patterns are shown to be a valid descriptor for an external dataset without this property, having a higher number of patients for later predictive analysis is likely to be beneficial for uncovering patterns. The author assumed that the benefits of the higher patient number would outweigh the negative effects from the randomized patients being included. This assumption was somewhat supported by the good generalization of the response patterns found in this experiment to the external validation dataset and further supported by the results from Experiment 3, where even with inclusion of the randomized patients, good prediction accuracy for the cluster structure was achieved.

After having shown mathematical validity and good generalization of the data set, the utility of the cluster structure can be investigated. This is different from attempts of stratifying patients into subgroups of patients with different underlying mechanisms. It is very possible that there are no “depression subgroups” that could be identified by a cluster structure on the level of symptom severity over time. In order to claim a possible mechanistic distinction between clusters, it would be necessary to test against the null hypothesis of there being no clusters and patients instead being described by a continuous distribution of some kind, which could be done by Bayesian modelling (see page 55).

In order to show utility, this testing doesn’t have to be performed. This is because the “real” structure underlying the identified clusters is unimportant, if a benefit – for clinical decision making in this case – can be shown. This can be conceptualized by imagining multiple normal gaussian distributions arbitrarily split into clusters. The most extreme clusters in this case would show distinctions between patients, regardless of the underlying real distribution in between. They thus have utility, without any information about the real distribution being required. See Figure 31 for a visualization of this concept.

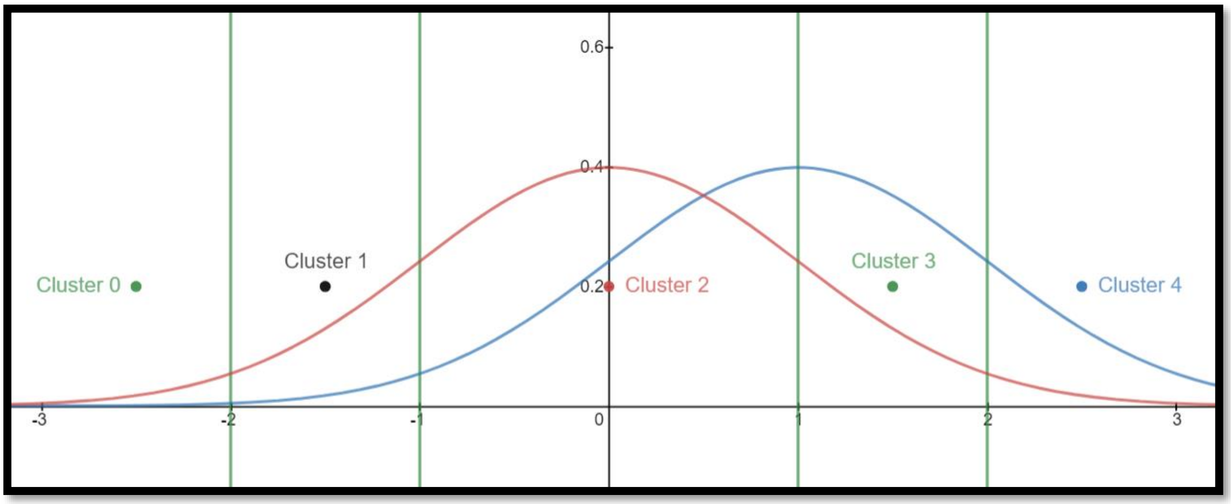


Figure 31: Two overlapping normal distributions being arbitrarily split into 5 clusters along the x axis. Clusters 0, 1 and 4 show good separation of the two hypothetical patient groups without any information about the real underlying distributions. This is an example of a cluster structure with utility. Graph created with [54].

Utility of the 5-cluster structure

Keeping the distinction between a claim of utility and a mechanistical claim in mind, the utility of the 5-cluster structure proposed in Experiment 1 should be examined in detail.

The between cluster differences in descriptive variables (see page 52) serve as a sanity check and first evidence of utility. As discussed in detail above, previously established predictors of response from clinical history as well as demographic data align with statistically significant differences between the clusters. While this result is unsurprising, given that the cluster structure can be interpreted in close relation to traditional response definitions (see page 46), establishing it was nonetheless important, because if non-correlation would have been shown, the feasibility of predicting cluster assignment early would have to undergo serious scrutiny.

Additional evidence of utility stems from comparing the cluster structure proposed here with proposed structures from previous literature (see page 50). The Delayed Improvement Cluster (2) runs roughly parallel to the first weeks of classes 7 and 8 from Uher et al. (2011), both of which show good overall response in the later course of treatment [10]. This suggests a subset of patients assigned to the Delayed Improvement Cluster (2) benefiting from a longer course of treatment, even though the traditional 50% response criterium would not be met after 4 weeks for these patients. This directly translates into a testable clinical hypothesis (formulated below) as per the goal of this thesis. The clinical information for this is based on the cluster assignment at week 4 alone, without any prediction being necessary. This establishes the possibility of the cluster structure being useful independent of any prediction task. To be considered evidence of clinical utility, further research to test the hypothesis is required, but deriving a testable hypothesis like this at least proves utility for future research.

In order to derive clinical utility, two possible sources of information were discussed in the thesis rationale (see page 16). An example of the first – being the early treatment response patterns over time – was discussed in the previous paragraph. For the second, treatment response needs to be predicted, instead of being classified ex post facto. This was done directly in Experiment 2, where treatment response as defined by traditional criteria was predicted based on clinical and demographic variables. This prediction was significantly better than random and performed better than the –

currently state of the art – early improvement criterium (see page 74 for detailed discussion of this). This shows – in addition to previous research on the subject (see page 11) – that predictive information that isn't being used by the current antidepressant treatment strategy or a strategy based on the early improvement criterium is contained in these clinical variables. In Experiment 3, the most promising variables for containing this information from Experiment 2 were used to predict (or classify based on partial information) later assignment to the 5-cluster structure. The high classification accuracies in conjunction with the high specificity of this prediction task before the traditional week 4 timepoint (see page 82 and 84 for detailed discussion into the interpretability of these metrics) act as proof for the conceptual value of combining the response time course with the predictive information.

With prediction of later cluster assignment being possible, the utility of the clusters can be derived from them benefitting from distinct clinical intervention. The most obvious hypothesis to derive from the cluster structure is then based on accelerating the decisions of current antidepressant treatment strategy forward for the most extreme group. The Non-Improvement (4) Cluster shows (on average) no treatment response after 4 weeks. If we can predict early, that a patient will belong to this cluster, it might be beneficial for this patient to receive an adapted course of treatment as early as possible. The utility of the cluster structure in this hypothesis over direct prediction of the traditional response criterium as in Experiment 2 stems from the focus on the extreme group of non-responders. This moves the decision boundary away from the edge cases of traditional response, which is directly associated with possible interventional cost. If traditional response is predicted, there will likely be edge-cases that would wrongly be predicted to be Non-Responders by week 4. These patients might then wrongly receive an early adaptation of treatment strategy when they might have benefitted from the longer treatment continuation that would've been chosen with traditional week 4 evaluation. When predicting the extreme group in the cluster structure, a patient wrongly predicted to be part of that group is still unlikely to show response by week 4 and thus would've received an adaptation of treatment strategy at that point anyway. This reduces the cost of the misclassification error. The cluster structure proposed here provides a data-derived, mathematically valid and generalizable (as shown in Experiment 1) way of defining a decision boundary for this task, which is likely preferable over arbitrarily adding a "margin of error" to the decision boundary of predicting traditional response. By combining this argument with the predictability of the cluster structure established in Experiment 3 and discussed in the previous paragraph, the utility of the cluster structure is supported further.

Formulating testable hypotheses

In the previous subsection, two possible clinical hypotheses were derived in the discussion of cluster utility. In this subsection these will be expanded upon and be formulated as testable hypotheses.

The first hypothesis was derived from the parallel response pattern of the Delayed Improvement (2) cluster to classes from Uher et al. (2011) which showed good treatment response in the later time course of treatment [10]. If the patients assigned to the Delayed Improvement (2) cluster have a comparatively large proportion of patients who show good antidepressant effects later in the treatment course, it might be beneficial to delay treatment evaluation for this group for longer than 4 weeks. This effect would likely be a trade-off between some patients receiving a delay of necessary treatment adaptation – meaning prolonged symptom severity and other patients not receiving unnecessary treatment adaptation which might lead to lower unwanted pharmacological effects and prevent ineffective treatment. Whether this trade-off is beneficial for a group on average will depend on the proportion of patients within this group.

Before this proportion can be investigated, the definition of the group warrants closer attention. It is not necessary, that defining the group for possible treatment delay by the cluster structure proposed in this paper is ideal. While the cluster structure was utilized to identify this group, the cluster interpretation along the traditional improvement and response criteria could result in preferable group proportion for the aforementioned trade-off. Though the inverse statement could be true just as likely, given the currently available data. Therefore, both definitions should be investigated.

In order to accurately assess the effects of a possible delay of treatment adaptation on the patient group, the overall proportion also isn’t the only factor. The time-duration of a possible delay would have strong effects of the cost-benefit analysis. It should therefore be established, which timepoint for evaluation would be ideal, given the proportions of the group.

With taking these points into account, the first hypothesis is formulated:

Hypothesis 1: A subgroup of patients with MDD benefits on average from non-adapted antidepressant treatment continuation until a delayed timepoint (later than week 4). This subgroup is defined either by assignment to the Delayed Improvement (2) cluster as proposed by this thesis or is defined as all patients without early improvement (> 20% improvement by week 2) who show no response (> 50% improvement by week 4) but do show delayed improvement (> 20% improvement by week 4). An optimal timepoint for delayed treatment adaptation in case of non-response to non-adapted treatment can be established for this subgroup.

The second hypothesis was derived from the predictability of the cluster structure. Since the data in this thesis suggests that assignment to the Non-Improvement (4) cluster can be predicted accurately for some patients before week 4, it seems obvious to accelerate treatment adaptation for these patients.

In order to formulate this hypothesis in a well-defined way, it is again important to consider the definition of the patient group an intervention would apply to. As opposed to Hypothesis 1, an intervention according to this hypothesis doesn’t necessarily have to be done for all patients for which corresponding cluster assignment is predicted. Instead, individual certainty of the prediction can be considered, since the random forest classifiers used for the prediction in Experiment 3 allow for estimation of a degree of certainty (see page 59). By considering the individual prediction certainty for each patient, a cost-benefit analysis doesn’t necessarily have to be performed groupwise but can be performed per-patient. This naturally then provides excellent support not only for clinical, but also for shared decision making – taking into account the individual patients preferences for safety or fast treatment adaptation.

If the algorithmic tools from this thesis are used for defining patients for possible early treatment adaptation, future investigation might be unnecessarily limited into these specific algorithmic tools. As discussed in detail above, the cluster structure proposed here might not necessarily be an ideal description of the underlying patient distribution (see page 86). Therefore, it might also not capture the optimal decision boundary for the prediction task (see page 88 for a detailed discussion of this). In addition, the prediction algorithm is just one possible algorithm limited to the set of clinical prediction variables available in the EMC dataset. This set of predictors is quite arbitrary and will most likely be incompatible with other research datasets or patient data from clinical practice.

Considering these points, the second hypothesis is formulated:

Hypothesis 2: Patients for which later assignment to an extreme group of Non-Improvers can be predicted with a high probability earlier than week 4 benefit from adaptation of treatment strategy at that earlier timepoint. Defining the extreme group as the Non-Improvement (4) cluster and using the prediction algorithm as proposed in this thesis is suitable and clinically beneficial over treatment strategy as usual.

Future research strategy regarding the hypotheses

The first steps for testing Hypothesis 1 should – in the authors opinion – be to describe the two possible patient group definitions regarding a possible benefit from non-adopted treatment continuation. For this, existing datasets in which patients received non-adapted antidepressant treatment for episodes longer than 4 weeks can be used. The patient groups according to both definitions can be identified in that dataset and classified by whether these patients express later response or remission. In a first step, it could then be time-stratified, which of the remitting patients would be detectable by treatment evaluation according to the response criterium at differing later timepoints – which patients of the subgroup show response by week 5, which by week 6 and so on. These proportions of responders found at different possible evaluation timepoints can then be combined with knowledge about antidepressant response after treatment adaptation (or if possible, with treatment adaptation data on the same patients from the same dataset). This combination can then be formulated as a mathematical optimization model (the details of which depend on the dataset) of the groupwise average cost-benefit analysis explained in the previous subsection. If this analysis, that can theoretically be performed purely in silico with already existing datasets, supports Hypothesis 1 and thus a benefit of treatment delay, it would constitute a strong scientific (and ethical) background for clinical trials to directly investigate the hypothesis. At the same time, this analysis is required to establish the optimal parameters (subgroup definition and treatment delay timepoint) for such clinical trials to take place. By these arguments, this proposed future strategy can fulfil both the necessary and sufficient conditions in order to move from the realm of psychiatric data science into clinical decision making, in case Hypothesis 1 isn’t falsified before then.

In regard to Hypothesis 2, the next research steps – in the authors opinion – should focus on validation and generalization of the prediction algorithm and the suitability of the cluster definition. Since any external datasets, with which this validation could be undertaken, likely have some difference in the set of available predictive variables and since practical clinical data is likely to differ even more in that regard, a more robust prediction algorithm that can – at least – handle missing variables should be trained. This could be as easy as retraining the prediction algorithm from Experiment 3 with blinding (setting to a constant for all patients) different combinations of variables, but depending on the available future validation datasets, other generalizations or imputations might be appropriate. This generalization of the prediction algorithm should ideally not be trained on the validation dataset at all and if required only be trained on part of the dataset, so that leak-free evaluation of prediction accuracy can occur without the limitations in this thesis (see page 84 for details). At the same time as the evaluation of prediction accuracy, the conformity of the – theoretical – early treatment adaptation after prediction with traditional treatment evaluation after 4 weeks can be tested, which should lead to direct estimates of numbers needed to treat and harm by the proposed intervention. Depending on these results, longitudinal clinical research can then be considered as the next step.

Additional utility of the developed algorithmic tools

In addition to the data-driven hypotheses derived from the cluster structure and the corresponding prediction algorithm, these algorithmic tools might have additional use for future scientific research.

Above, the difference of a clustering with mechanistic claim and a utility clustering as done in this thesis was discussed in detail (see page 86). What was not explored in more detail there, is that a utility-based clustering can be used for further investigation into mechanistic hypotheses without originally having this claim in mind. This is due to the effect illustrated in Figure 31, where an arbitrary clustering can lead to good separation between overlapping distributions without assumptions of the underlying distributions. The clustering proposed in this thesis for example provides a mathematically optimized and generalizable way, to separate the extreme groups of Non-Responders (4) and Early Responders (0). By calculation of the silhouette sample score, the algorithm also provides a way to select the most typical patients for these groups. This might be beneficial over just identifying the most extreme cases, since these might also be quite untypical. The associative comparison of these extreme groups – for example in the field of genetics or epigenetics – could then possibly uncover group-level associations, which might be masked while not investigating extreme groups or when selecting atypical patients for further investigation. In case such research is undertaken, the results would need to be verified on patients not part of the initial extreme group selection, otherwise a mechanistic claim of the cluster structure would be assumed without this being justifiable. This example of extreme groups can obviously be generalized to the remaining cluster structure. If research is undertaken into the difference or commonalities between different trajectories of the early treatment response time-course, the cluster structure and its algorithm proposed here do provide a way of defining and separating groups of interest and assessing typicality of a given patient for that group.

The predictive algorithm built in Experiment 2 was discussed in comparison to the early improvement criterium above (see page 74). The early improvement criterium can be seen as the current “state of the art” in early prediction of antidepressant response (see page 10). Showing, that the state of the art can be outperformed by a rather simple prediction algorithm without feature selection (which wasn’t done until Experiment 3, so all cross-validation accuracies in Experiment 2 are interpretable leak-free), is a good proof of concept for the usage of such prediction algorithms in clinical decision making in favour of the early improvement criterium. Before prediction algorithms like this can be deployed in clinical practice, they need to be designed robust to changing sets of predictive variables and missing values, which are both characteristics to be expected from practical clinical data. This likely requires future work in combining multiple research datasets and collecting an extensive pool of data from clinical practice. The success of the direct prediction algorithm developed in this thesis provides support for the possible value of such research, which is significant scientific utility as additional result of Experiment 2.

The prediction algorithms from Experiments 2 and 3 have an additional use cases for science. Because they selected a set of informative clinical variables for prediction, any potentially predictive variables that are added to this set can be evaluated based on comparison of predictive performance. This is especially relevant in the field of biomarker search, since testing for increase of predictive performance over clinical variables alone can give a strong argument for validity for any biomarkers. If biomarkers do not increase predictive performance, the biomarker might of course still be interesting to better understand the pathophysiology of depression or as on objectifiable (forensic) marker of depression, but it will then unlikely be of use in clinical practice – the information contained within the biomarker can theoretically be gained just as well by clinical observation alone, which will usually be much cheaper.

Utilization of the predictive algorithms in this manner is currently already being planned for future research in the group for which this thesis was written, which emphasizes the potential value of the algorithms for this purpose.

Conclusion

MDD is disease with large importance both on the individual patient level and the healthcare system level. To improve treatment response rates and to mitigate the multitude of adverse effects, research into the most effective treatment strategies is paramount. In order to improve upon current treatment strategy, research into the underlying pathophysiology of depression, into novel pharmacological substances and into specialised psychotherapeutical interventions (and into many other areas) are being undertaken. But while these fields are beginning to benefit clinical practice, it is also important to make the best use of existing information and existing treatment strategies.

The current antidepressant treatment strategy is inadequate in that regard. The goal of this thesis was to investigate sources of clinical information that were potentially underused and search for possible hypotheses regarding clinical decision making based on this information. And while hypotheses were found by utilizing the machine learning methodology in this thesis, this is only a small step in establishing better utilization of the existing clinical data.

Continued investigation into the hypotheses from this thesis as well as into other potential sources of underused information has the potential to improve practical clinical decision making and consecutively diminish patient suffering and societal impact. And until the aforementioned other areas of research bring big leaps into practical treatment of depression, the small incremental improvements to be expected from this continued investigation are one of only a few potential ways of providing an immediate benefit to suffering patients.

Abstract

Current antidepressant treatment strategies in MDD evaluate treatment response only after 4 weeks of treatment duration. The early improvement criterium (20% sum score improvement after 2 weeks), which is currently state of the art for earlier prediction of treatment response, wasn’t proven to be an effective trigger for clinical decision making so far. This thesis investigates potential sources of predictive information other than symptom severity sum scores, in order to find new hypotheses for future treatment strategies. In Experiment 1, early treatment response patterns over time are identified by clustering with the k-means-algorithm and several possible cluster-structures are evaluated for mathematical fit as well as clinical interpretability. A structure with 5 clusters of early response is identified as a candidate for further investigation and hypothesis building. In Experiment 2, traditional clinical response and remission criteria are predicted using random forest classifiers with different sets of clinical variables at different timepoints as predictors. The classifiers are evaluated in comparison to the early improvement criterium which is being outperformed for some of the predictor sets at any timepoint. This shows that predictive information is contained in clinical variables other than the sum score. These variables are assessed and selected for further model building based on their relative feature importance scores. In Experiment 3, a random forest classifier based on the variables selected in Experiment 2 is trained to predict assignment to the clusters from Experiment 1, thereby combining the two sources of predictive information. The results show this prediction to be possible above the zero-information rate for later timepoints. In the combined discussion, the results from these three Experiments are combined to formulate two new hypotheses for treatment strategy. The first hypothesis assumes that the “Delayed Improvement” cluster from Experiment 1 benefits (on average) from treatment continuation longer than 4 weeks and the second hypothesis assumes that patients that will likely – based on predictions like in Experiment 3 – be part of the “Non Improvement” cluster from Experiment 1 benefit from early medication change. The role of the algorithms from this thesis for research into the hypotheses as well as their additional scientific use is discussed.

Zusammenfassung

Aktuelle antidepressive Therapiestrategien für die Behandlung depressiver Erkrankungen evaluieren Therapieansprechen nach einer Therapiedauer von 4 Wochen. Das sogenannte „early improvement“-Kriterium (20% Symptom-Summenscore Verbesserung in 2 Wochen nach Therapiebeginn), der aktuelle Stand der Wissenschaft hinsichtlich früherer Verlaufsprädiktion, konnte nach bisheriger Datenlage bislang nicht als verlässliche klinische Entscheidungshilfe etabliert werden. Diese Dissertation untersucht potenzielle Quellen für neue prädiktive Information, um überprüfbare Hypothesen für zukünftige Therapiestrategien zu generieren. In Experiment 1 wurden typische Verlaufsmuster klinischen Ansprechens identifiziert, indem Therapieverläufe mit dem k-means-Algorithmus gruppiert werden. Mehrere mögliche Gruppierungen wurden hinsichtlich der Güte der mathematischen Beschreibung sowie ihrer klinischen Interpretierbarkeit untersucht und diskutiert. Eine Gruppierung in 5 typische Verlaufsmuster wurde als Kandidat für weitere Untersuchungen sowie Hypothesenbildung identifiziert. In Experiment 2 wurden random forest Klassifikationsalgorithmen eingesetzt, um Eintreten der traditionellen Response- und Remissionskriterien der Depression zu prädizieren. Hierzu wurden mehrere Gruppen von Variablen zu unterschiedlichen Zeitpunkten im Behandlungsverlauf als Prädiktoren eingesetzt. Die resultierenden Klassifikatoren werden mit dem „early improvement“ Kriterium verglichen, dabei erreichten erstere an jedem Zeitpunkt eine höhere Genauigkeit für wenigstens einen Teil der Prädiktorengruppen. Dies zeigte den prädiktiven Wert klinischer Variablen über die Summenscores hinaus. Die prädiktiven Variablen wurden anhand ihrer Bedeutung für die Klassifikation untersucht und für weitere Modellbildung ausgewählt. In Experiment 3 wurde ein random forest Klassifikationsalgorithmus mit den in Experiment 2 gewählten Prädiktionsvariablen trainiert, um Zugehörigkeit zu den Verlaufsgruppen aus Experiment 1 zu prädizieren. So wurden die beiden potenziellen Quellen für klinischen Informationsgewinn kombiniert. Es wurde gezeigt, dass diese Prädiktion im späteren Therapieverlauf über die Null-Informationsrate hinaus möglich ist. In der Diskussion wurden die Ergebnisse der drei Experimente zusammengefügt, um zwei neue Hypothesen zur antidepressiven Therapiestrategie abzuleiten. Die erste Hypothese nimmt an, dass Patienten aus der „Delayed Improvement“ Gruppe aus Experiment 1 im Durchschnitt von einer längeren antidepressiven Therapiedauer als 4 Wochen profitieren. Die zweite Hypothese nimmt einen Nutzen eines frühen

Medikationswechsels bei Patienten an, bei denen eine spätere Zuordnung zur „Non Improvement“ Gruppe aus Experiment 1 mit hoher Wahrscheinlichkeit prädiziert wird (vergleichbar mit der Prädiktion in Experiment 3). Die Rolle der Algorithmen aus dieser Dissertation und deren Rolle für die zukünftige Hypothesenprüfung sowie zusätzliche wissenschaftliche Nutzung wurden diskutiert.

References

[1] R. C. Kessler, P. Berglund, O. Demler, R. Jin, D. Koretz, K. R. Merikangas, J. A. Rush, E. E. Walters und P. S. Wang, „The Epidemiology of Major Depressive Disorder - Results From the National Comorbidity Survey Replication (NCS-R),“ *Journal of the American Medical Association,* pp. 289(23):3095-3105, 18 06 2003.

[2] F. Jacobi, H.-U. Wittchen, C. Hölting, M. Höfler, H. Pfister, N. Müller und R. Lieb, „Prevalence, co-morbidity and correlates of mental disorders in the general population: results from the German Health Interview and Examination Survey (GHS),“ *Psychological Medicine,* pp. 597-611.

[3] World Health Organization, „Depression and Other Common Mental Disorders: Global Health Estimates,“ World Health Organization, Geneva, 2017.

[4] P. E. Greenberg, A.-A. Fournier, T. Sisitsky, C. T. Pike und R. C. Kessler, „The Economic Burden of Adults With Major Depressive Disorder in the United States (2005 and 2010),“ *Journal of Clinical Psychiatry,* p. 76:2, 2015.

[5] J. R. Randall, R. Walld, G. Finlayson, J. Sareen, P. J. Martens und J. M. Bolton, „Acute Risk of Suicide and Suicide Attempts Associated With Recent Diagnosis of Mental Disorders: A Population-Based, Propensity Score–Matched Analysis,“ *The Canadian Journal of Psychiatry,* pp. 59(10), 531–538, 2014.

[6] DGPPN, BÄK, KBV, AWMF (Hrsg.) für die Leitliniengruppe Unipolare Depression, S3-Leitlinie/Nationale VersorgungsLeitlinie Unipolare Depression – Langfassung, 2. Auflage. Version 5, 2015.

[7] M. Bauer, A. Pfennig, E. Severus, P. C. Whybrow, J. Angst und H.-J. Möller, „World Federation of Societies of Biological Psychiatry (WFSBP) Guidelines for Biological Treatment of Unipolar Depressive Disorders, Part 1: Update 2013 on the acute and continuation treatment of unipolar depressive disorders,“ *The World Journal of Biological Psychiatry,* p. 14: 334–385, 2013.

[8] A. Cipriani, T. A. Furukawa, G. Salanti, A. Chaimani, L. Z. Atkinson, Y. Ogawa, S. Leucht, H. G. Ruhe, E. H. Turner, J. P. T. Higgins, M. Egger, N. Takeshima, Y. Hayasaka, H. Imaj, K. Shinohara, A. Tajika, J. P. A. Ioannidis und J. R. Geddes, „Comparative efficacy and acceptability of 21 antidepressant drugs for the acute treatment of adults with major depressive disorder: a systematic review and network meta-analysis," *The Lancet,* pp. 391:1357-1366, 7 April 2018.

[9] S. Wagner, A. Engel, J. Engelmann, D. Herzog, N. Dreimüller, M. B. Müller, A. Tadic und K. Lieb, „Early improvement as a resilience signal predicting later remission to antidepressant treatment in patients with Major Depressive Disorder: Systematic review and meta-analysis," *Journal of Psychiatric Research,* pp. 94:96-106, 2017.

[10] R. Uher, O. Mors, M. Rietschel, A. Rajewska-Rager, A. Petrovic, A. Zobel, N. Henigsberg, J. Mendlewicz, K. J. Aitchison, A. Farmer und P. McGuffin, „Early and delayed onset of response to antidepressants in individual trajectories of change during treatment of major depression: A secondary analysis of data from the Genome-Based Therapeutic Drugs for Depression (GENDEP) Study.," *Journal of Clinical Psychiatry,* pp. 72(11):1478-1484, 2011.

[11] A. Tadic, D. Wachtlin, M. Berger, D. F. Braus, D. van Calker, N. Dahmen, N. Dreimüller, A. Engel, S. Gorbulev, I. Helmreich, A.-K. Kaiser, K. Kronfeld, K. F. Schlicht, O. Tüscher, S. Wagner, C. Hiemke und K. Lieb, „Randomized controlled study of early medication change for non-improvers to antidepressant therapy in major depression – The EMC trial," *European Neuropsychopharmacology,* pp. 26: 705-716, 2016.

[12] A. Tadić, S. Gorbulev, N. Dahmen, C. Hiemke, F. D. Braus, J. Röschke, D. van Calker, D. Wachtlin, K. Kronfeld, T. Gorbauch, M. Seibert-Grafe und K. Lieb, „Rationale and design of the randomised clinical trial comparing early medication change (EMC) strategy with treatment as usual (TAU) in patients with Major Depressive Disorder - the EMC trial," *Trials,* p. 11:21, 2010.

[13] American Psychiatric Association, Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, Washington DC: APA, 2013.

[14] World Health Organization, The ICD-10 Classification of Mental and Behavioural Disorders. Clinical Descriptions and Diagnostic Guidelines., Geneva: World Health Organization, 1992.

[15] E. I. Fried und R. M. Nesse, „Depression sum-scores don't add up: why analyzing specific depression symptoms is essential,“ *BMC Medicine,* p. 13:72, 2015.

[16] M. Hamilton, „A RATING SCALE FOR DEPRESSION,“ *Journal of Neurology, Neurosurgery and Psychiatry,* pp. 23:56-62, 1960.

[17] S. A. Montgomery und M. Åsberg, „A New Depression Scale Designed to be Sensitive to Change,“ *The British Journal of Psychiatry,* pp. 134: 382-389, 1979.

[18] A. T. Beck, R. A. Steer und G. K. Brown, Manual for the Beck Depression Inventory-II, San Antonia, TX: Psychological Corporation, 1996.

[19] A. J. Rush, C. M. Gullion, M. R. Basco, R. B. Jarrett und M. H. Trivedi, „The Inventory of Depressive Symptomatology (IDS): psychometric properties,“ *Psychological Medicine,* pp. 26: 477-486, 1996.

[20] E. Frank, R. F. Prien, R. B. Jarrett, M. B. Keller, D. J. Kupfer, P. W. Lavori, J. A. Rush und M. M. Weissman, „Conceptualization and Rationale for Consensus Definitions of Terms in Major Depressive Disorder - Remission, Recovery, Relapse, and Recurrence,“ *Archives of General Psychiatry,* pp. 48(9):851-855, 1991.

[21] K. Lieb, S. Frauenknecht, S. Brunnhuber und C. Wewetzer, Intensivkurs Psychiatrie und Psychotherapie, München: Elsevier, 2016.

[22] A. A. Nierenberg, A. H. Farabaugh, J. E. Alpert, J. Gordon, J. J. Wothington, J. F. Rosenbaum und M. Fava, „Timing of Onset of Antidepressant Response With Fluoxetine Treatment,“ *American Journal of Psychiatry,* p. 157:1423–1428, 2000.

[23] A. Szegedi, M. J. Müller, I. Anghelescu, C. Klawe, R. Kohnen und O. Benkert, „Early Improvement Under Mirtazapine and Paroxetine Predicts Later Stable Response and Remission With High Sensitivity in Patients With Major Depression,“ *The Journal of clinical psychiatry,* pp. 64(4): 413-420, 2003.

[24] M. M. Katz, J. L. Tekell, C. L. Bowden, S. Brannan, J. P. Houstonn, N. Berman und A. Frazer, „Onset and Early Behavioral Effects of Pharmacologically Different Antidepressants and Placebo in Depression," *Nature Neuropsychopharmacology,* p. 29: 566–579, 2004.

[25] A. Szegedi, W. T. Jansen, A. P. P. van Willenburg, E. van der Meulen, H. H. Stassen und M. E. Thase, „Early Improvement in the First 2 Weeks as a Predictor of Treatment Outcome in Patients With Major Depressive Disorder: A Meta-Analysis Including 6562 Patients," *The Journal of Clinical Psychiatry,* pp. 70(3): 344-353, 2009.

[26] S. Nakajima, H. Uchida, T. Suzuki, K. Watanabe, J. Hirano, T. Yagihashi, H. Takeuchi, T. Abe, H. Kashima und M. Mimura, „Is switching antidepressants following early nonresponse more beneficial in acute-phase treatment of depression?: A randomized open-label trial," *Progress in Neuro-Psychopharmacology & Biological Psychiatry,* pp. 35: 1983-1989, 2011.

[27] M. Fava, J. A. Rush, M. H. Trivedi, A. A. Nierenberg, M. E. Thase, H. A. Sackeim, F. M. Quitkin, S. Wisniewski, P. W. Lavori, J. F. Rosenbaum, D. J. Kupfer und for the STAR∗D Investigators Group, „Background and rationale for the Sequenced Treatment Alternatives to Relieve Depression (STAR∗D) study," *Psychiatric Clinics of North America,* pp. 26:457-494, 2003.

[28] J. A. Rush, S. R. Wisniewski, D. Warden, J. F. Luther, L. L. Davis, M. Fava, A. A. Nierenberg und M. H. Trivedi, „Selecting Among Second-Step Antidepressant Medication Monotherapies - Predictive Value of Clinical, Demographic, or First-Step Treatment Features," *Archives of General Psychiatry,* p. 65(8): 870, 2008.

[29] A. F. Kozel, M. S. Madhukar, H. Trivedi, S. R. Wisniewski, S. Miyahara, M. M. Husain, M. Fava, B. Lebowitz, S. Zisook und J. A. Rush, „Treatment Outcomes for Older Depressed Patients With Earlier Versus Late Onset of First Depressive Episode," *The American Journal of Geriatric Psychiatry,* pp. 16 (1):58-64, 2008.

[30] A. Drago und A. Serretti, „Sociodemographic Features Predict Antidepressant Trajectories of Response in Diverse Antidepressant Pharmacotreatment Environments.," *Journal of Clinical Psychopharmacology,* p. 31(3): 345–348., 2011.

[31] A. M. Chekroud, R. J. Zotti, Z. Shehzad, R. Gueorguieva, M. K. Johnson, M. H. Trivedi, T. D. Cannon, J. H. Krystal und P. R. Corlett, „Cross-trial prediction of treatment outcome in depression: a machine learning approach," *The Lancet Psychiatry,* pp. 3 (3): 243-250, 2016.

[32] J. A. Rush, M. H. Trivedi, J. W. Stewart, A. A. Nierenberg, M. Fava, B. T. Kurian, D. Warden, D. W. Morris, J. F. Luther, M. M. Husain, I. A. Cook, R. C. Shelton, I. M. Lesser, S. G. Kornstein und S. R. Wisniewski, „Combining Medications to Enhance Depression Outcomes (CO-MED): Acute and Long-Term Outcomes of a Single-Blind Randomized Study," *American Journal of Psychiatry,* p. 168: 689–701, 2011.

[33] R. Paul, T. F. M. Andlauer, D. Czamara, D. Hoehn, S. Lucae, B. Pütz, C. M. Lewis, R. Uher, B. Müller-Myhsok, M. Ising und P. G. Sämann, „Treatment response classes in major depressive disorder identified by model-based clustering and validated by clinical prediction models," *Translational Psychiatry,* p. 9:187, 2019.

[34] J. M. Hennings, T. Owashi, E. B. Binder, S. Horstmann, A. Menke, S. Kloiber, T. Dose, B. Wollweber, D. Spieler, T. Messer, R. Lutz, H. Künzel, T. Bierner, T. Pollmächer, H. Pfister, T. Nickel, A. Sonntag, M. Uhr, M. Ising, F. Holsboer und S. Lucae, „Clinical characteristics and treatment outcome in a representative sample of depressed inpatients – Findings from the Munich Antidepressant Response Signature (MARS) project," *Journal of Psychiatric Research,* pp. 43: 215-229, 2009.

[35] R. Uher, N. Perroud, M. Y. Ng, J. Hauser, N. Henigsberg, W. Maier, O. Mors, A. Placentino, M. Rietschel, D. Souery, T. Zagar, P. M. Czerski, B. Jerman, E. R. Larsen, T. G. Schulze, A. Zobel, S. Cohen-Woods, K. Pirlo, A. W. Butler, P. Muglia, M. R. Barnes, M. Lathrop, A. Farmer, G. Breen, K. J. Aitchison, I. Craig, C. M. Lewis und P. McGuffin, „Genome-Wide Pharmacogenetics of Antidepressant Response in the GENDEP Project," *The American Journal of Psychiatry,* p. 167: 555–564, 2010.

[36] D. V. Sheehan, Y. Lecrubier, H. K. Sheehan, P. Amorim, J. Janvas, E. Weiller, H. Thierry, R. Baker und G. C. Dunbar, „The Mini-International Neuropsychiatric Interview (M.I.N.I.): The Development and Validation of a Structured Diagnostic

Psychiatric Interview for DSM-IV and ICD-10," *Journal of clinical Psychiatry,* pp. 59[suppl 20]:22-23, 1998.

[37] M. B. First, M. Gibbon, R. L. Spitzer, J. B. W. Spitzer und L. S. Benjamin, SCID-II: Structured Clinical Interview for DSM-IV Axis II Personality disorders: User's Guide., New York, NY: American Psychiatric Press, 1997.

[38] G. van Rossum, Python tutorial, Technical Report CS-R9526, Amsterdam: Centrum voor Wiskunde en Informatica (CWI), 1995.

[39] W. McKinney, „Data Structures for Statistical Computing in Python," *Proceedings of the 9th Python in Science Conference,* pp. 51-56, 2010.

[40] T. E. Oliphant, A guide to NumPy, USA: Trelgol Publishing, 2006.

[41] J. D. Hunter, „Matplotlib: A 2D Graphics Environment," *Computing in Science & Engineering,* pp. 9: 90-95, 2007.

[42] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot und E. Duchesnay, „Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research,* pp. 12: 2825-2830, 2011.

[43] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, J. K. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. Carey, I. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt und and Scipy 1.0 Contributors, „SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods,* pp. 17: 261--272, 2020.

[44] E. W. Weissstein, „K-Means Clustering Algorithm.," [Online]. Available: http://mathworld.wolfram.com/K-MeansClusteringAlgorithm.html. [Zugriff am 16 01 2019].

[45] D. J. Ketchen Jr. und C. L. Shook, „THE APPLICATION OF CLUSTER ANALYSIS IN STRATEGIC MANAGEMENT RESEARCH: AN ANALYSIS AND CRITIQUE,“ *Strategic Management Journal,* pp. Vol. 17, 441 -458, 1996.

[46] P. J. Rousseeuw, „Silhouettes: a graphical aid to the interpretation and validation of cluster analysis,“ *Journal of Computational and Applied Mathematics ,* pp. Vol 20, 53-65, 1987.

[47] A. Rohatgi, „WebPlotDigitizer Version 4.2,“ [Online]. Available: https://automeris.io/WebPlotDigitizer. [Zugriff am 01 07 2019].

[48] R. Suganya und R. Shanthi, „Fuzzy C- Means Algorithm- A Review,“ *International Journal of Scientific and Research Publications,* pp. Volume 2, Issue 11, 2012 .

[49] S. Wade und Z. Ghahramani, „Bayesian Cluster Analysis: Point Estimation and Credible Balls (with Discussion),“ *Bayesian Analysis,* pp. 13, Number 2, pp. 559–626, 2018.

[50] R. D. Peng, „Hierarchical Clustering,“ in *Exploratory Data Analysis with R*, leanpub.com, 2016, p. Chapter 12.

[51] T. Ganegedara, „Light on Math Machine Learning: Intuitive Guide to Understanding Decision Trees,“ Towards Data Science, 29 11 2018. [Online]. Available: https://towardsdatascience.com/light-on-math-machine-learning-intuitive-guide-to-understanding-decision-trees-adb2165ccab7. [Zugriff am 20 01 2019].

[52] R. Eulogio, „Introduction to Random Forests,“ 12 08 2017. [Online]. Available: https://www.datascience.com/resources/notebooks/random-forest-intro. [Zugriff am 18 07 2019].

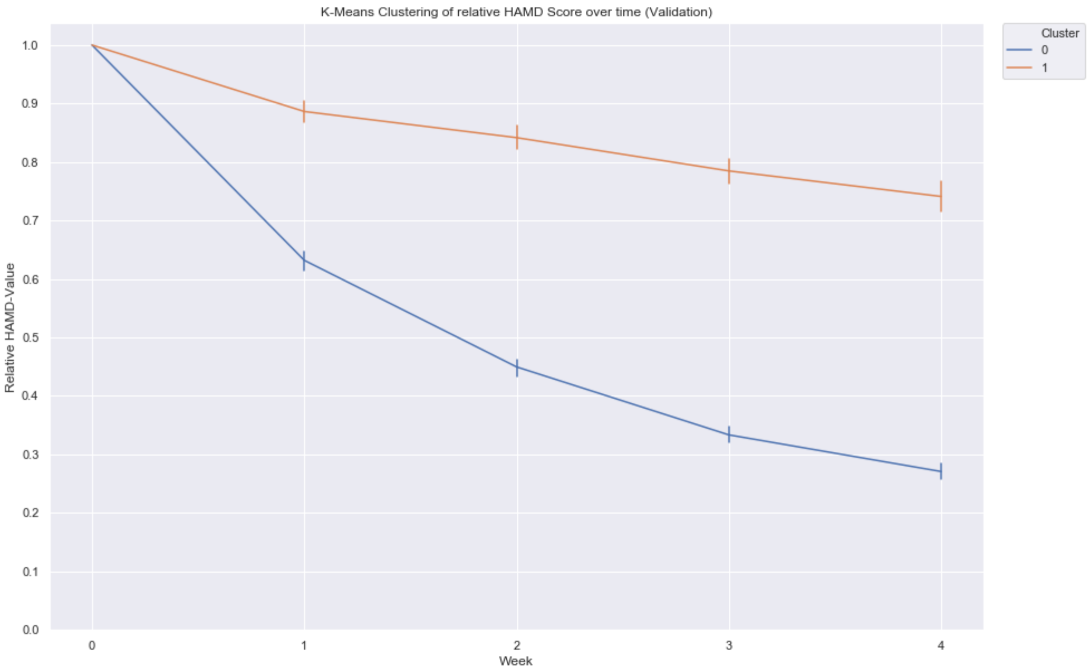
[53] G. Jansen, „Quora - Should inputs to random forests be normalized,“ Quora.com, 12 September 2012. [Online]. Available: https://www.quora.com/Should-inputs-to-random-forests-be-normalized. [Zugriff am 28 August 2019].

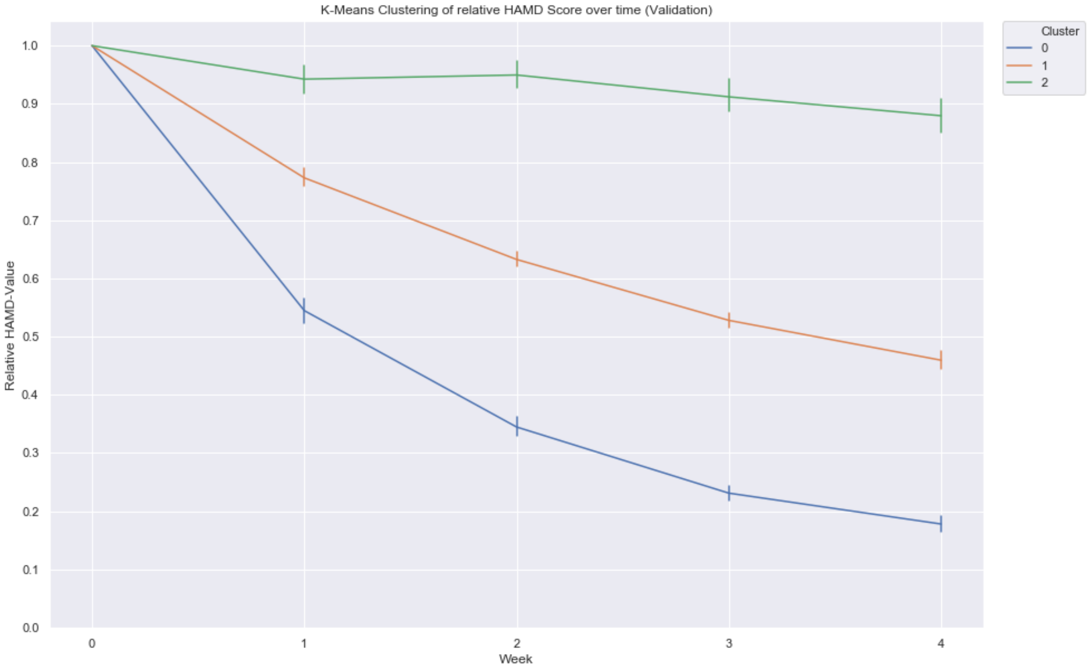
[54] Desmos, Inc., „Desmos.com,“ 2019. [Online]. Available: https://www.desmos.com/.

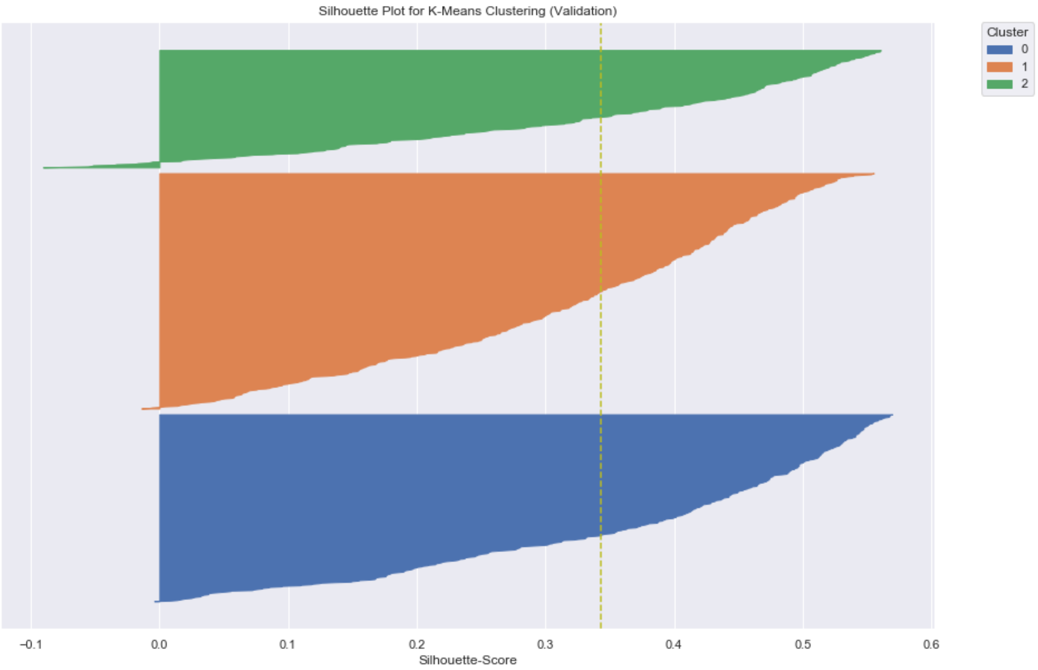
[55] R. Machado-Vieira, J. Baumann, C. Wheeler-Castillo, D. Latov, I. D. Henter, G. Salvadore und C. A. Zarate Jr., „The Timing of Antidepressant Effects: A Comparison of Diverse Pharmacological and Somatic Treatments,“ *Pharmaceuticals,* pp. 3:19-41, 2010.

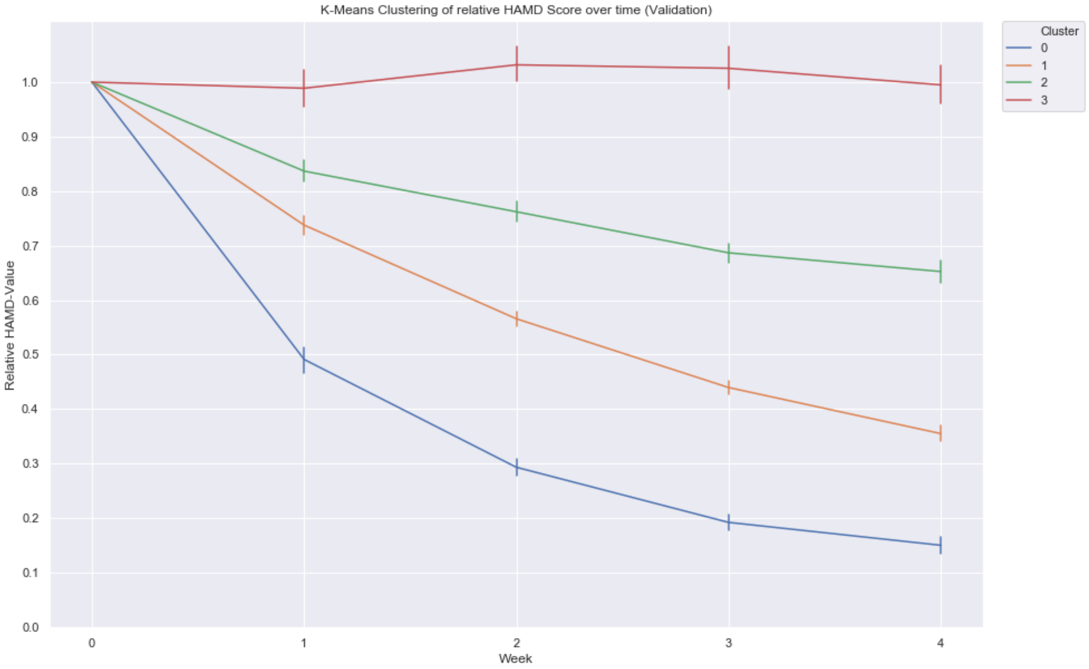
Appendix

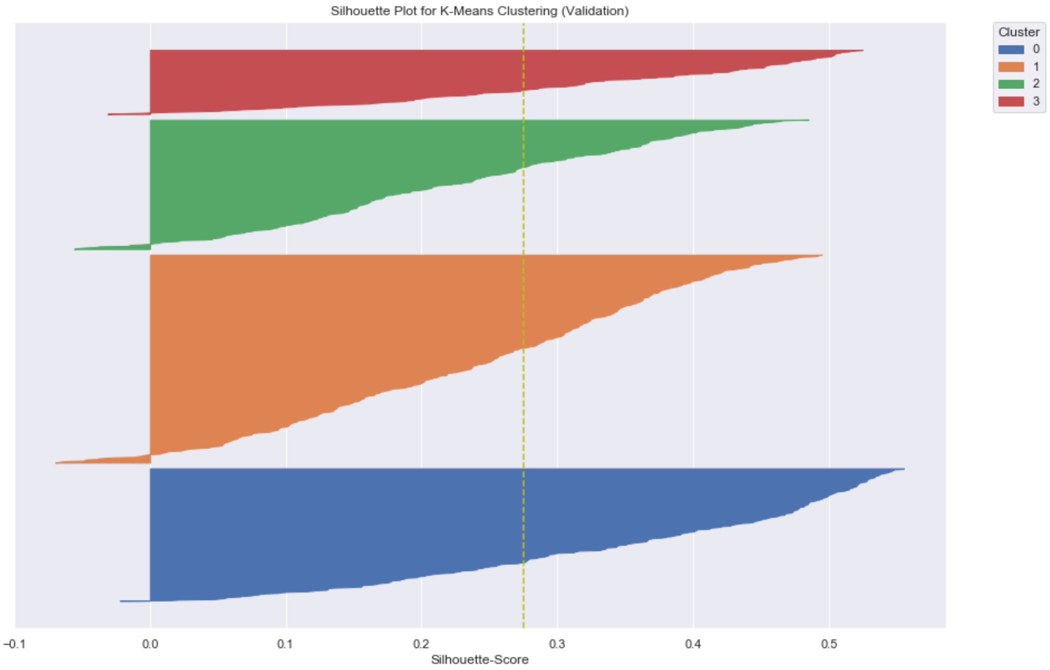
Experiment 1

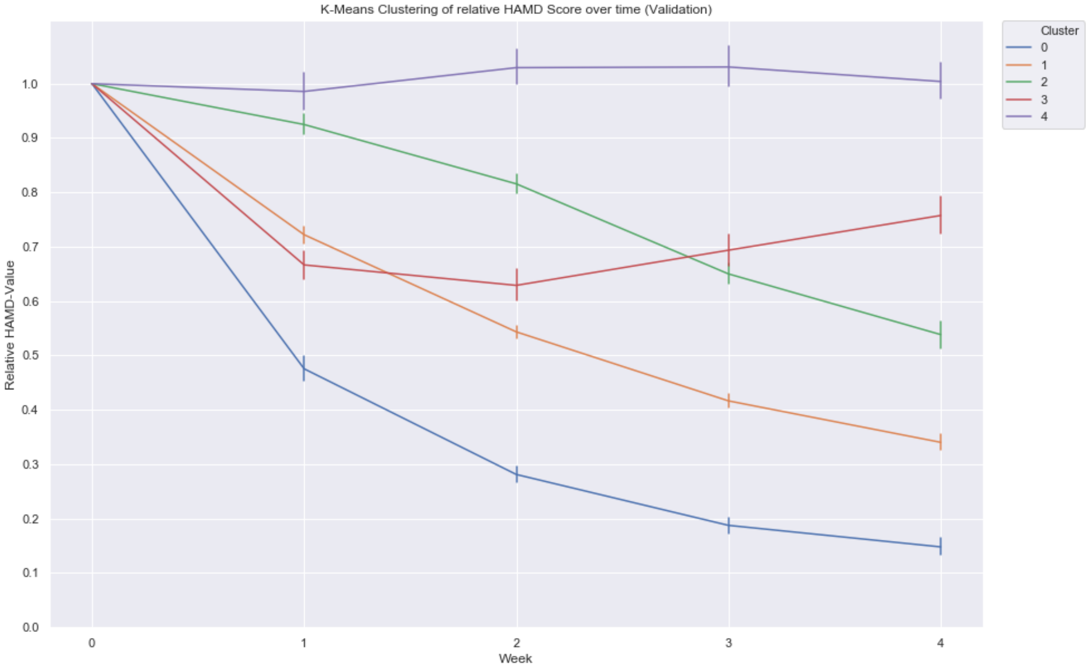


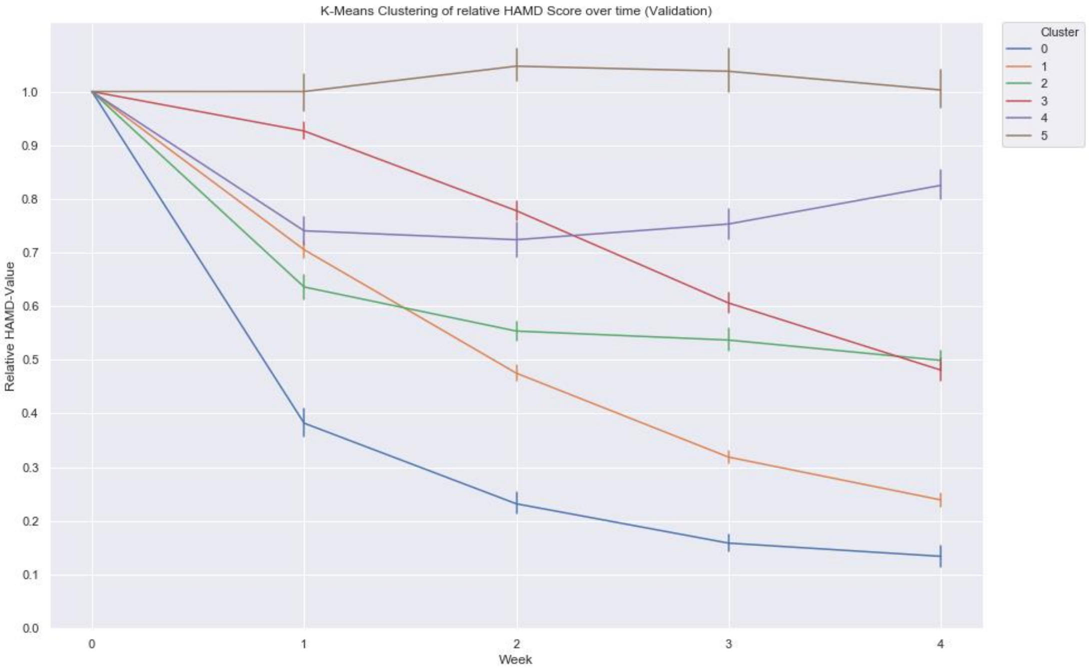


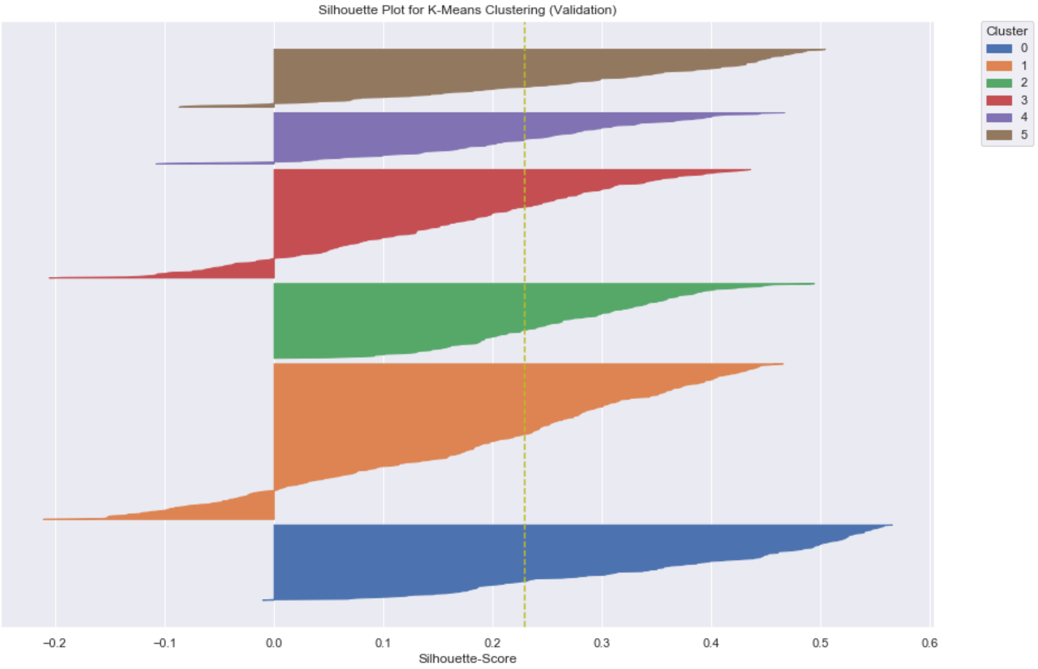

Silhouette Plot for K-Means Clustering (Validation)
Cluster
0
1
2
−0.1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Silhouette-Score

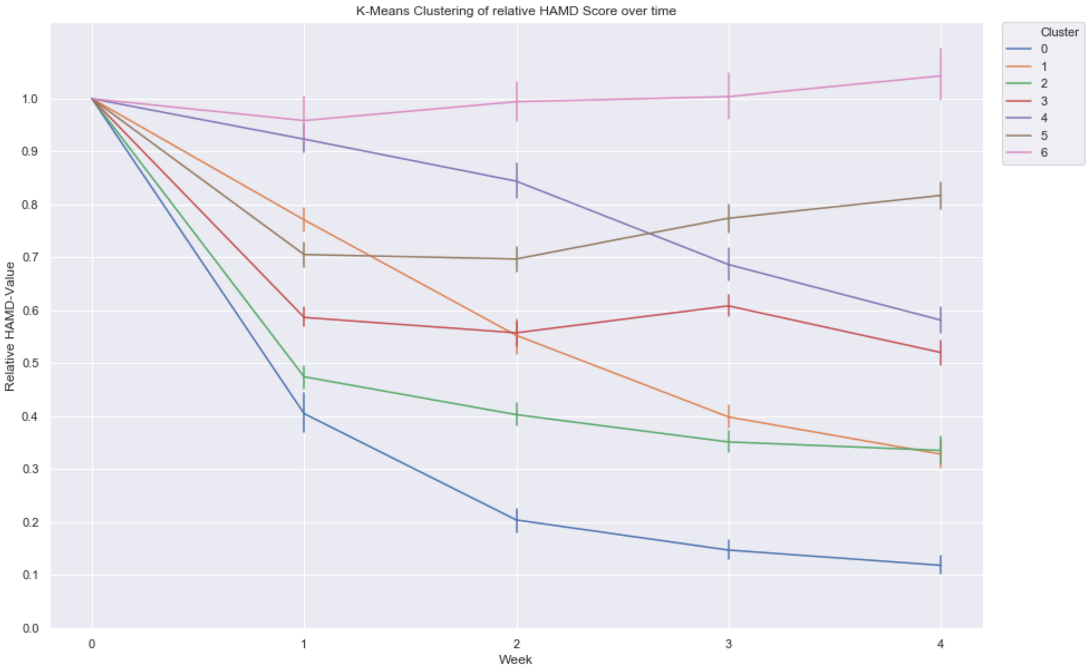

K-Means Clustering of relative HAMD Score over time (Validation)
Cluster
0
1
2
3
Relative HAMD-Value
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
1
2
3
4
Week

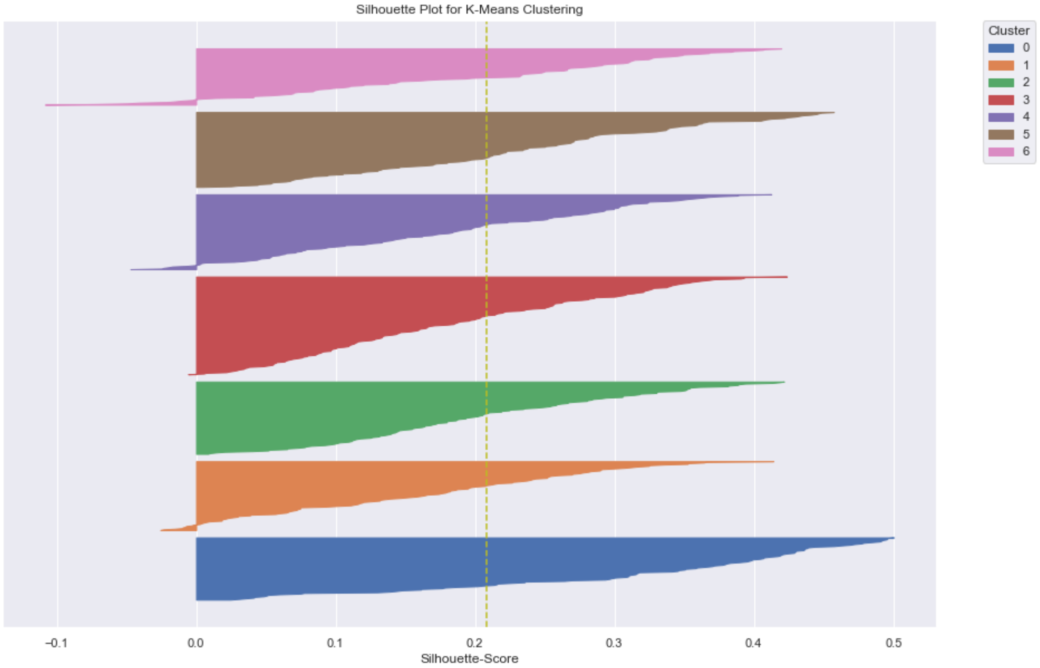

Silhouette Plot for K-Means Clustering (Validation)
Cluster
0
1
2
3
−0.1
0.0
0.1
0.2
0.3
0.4
0.5
Silhouette-Score

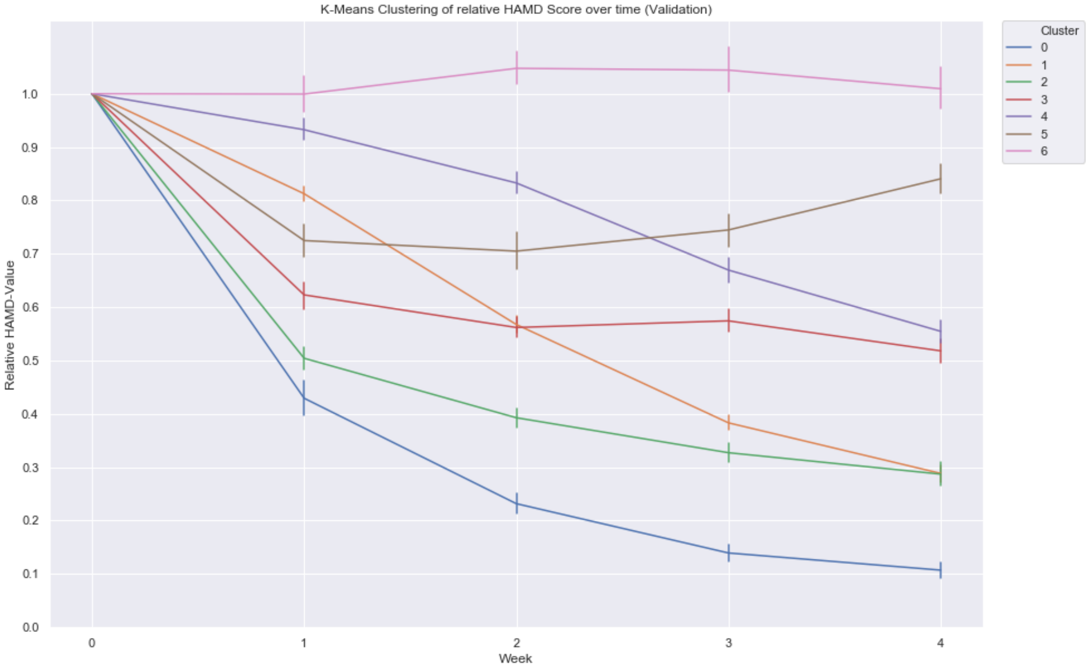

K-Means Clustering of relative HAMD Score over time (Validation)
Cluster
0
1
2
3
4
Relative HAMD-Value
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
1
2
3
4
Week

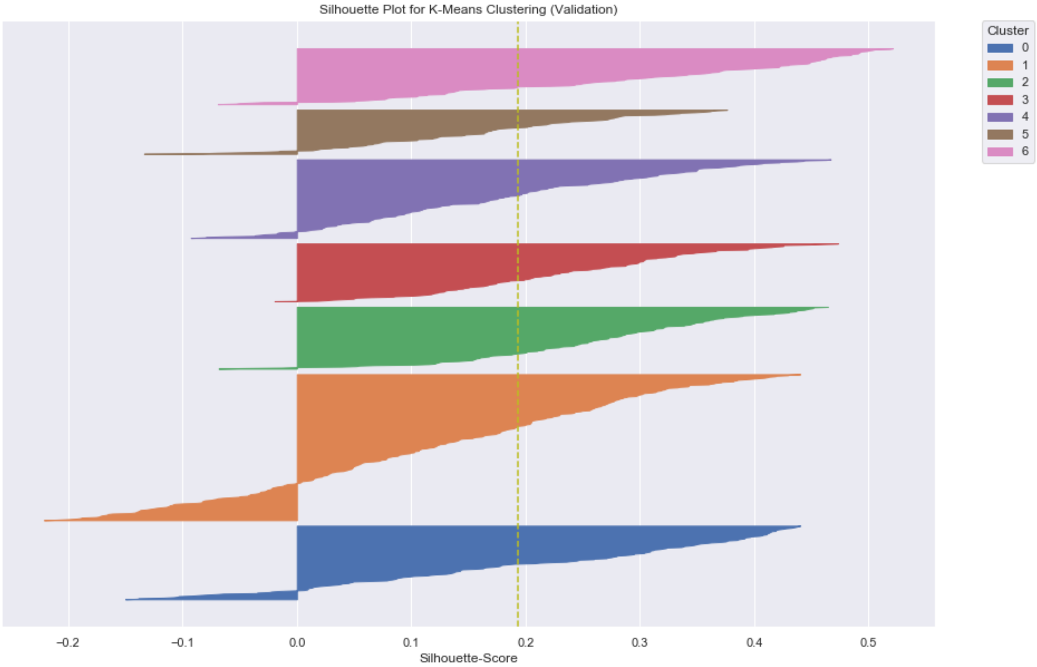

K-Means Clustering of relative HAMD Score over time (Validation)
Relative HAMD-Value
Week
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
0
1
2
3
4
Cluster
0
1
2
3
4
5

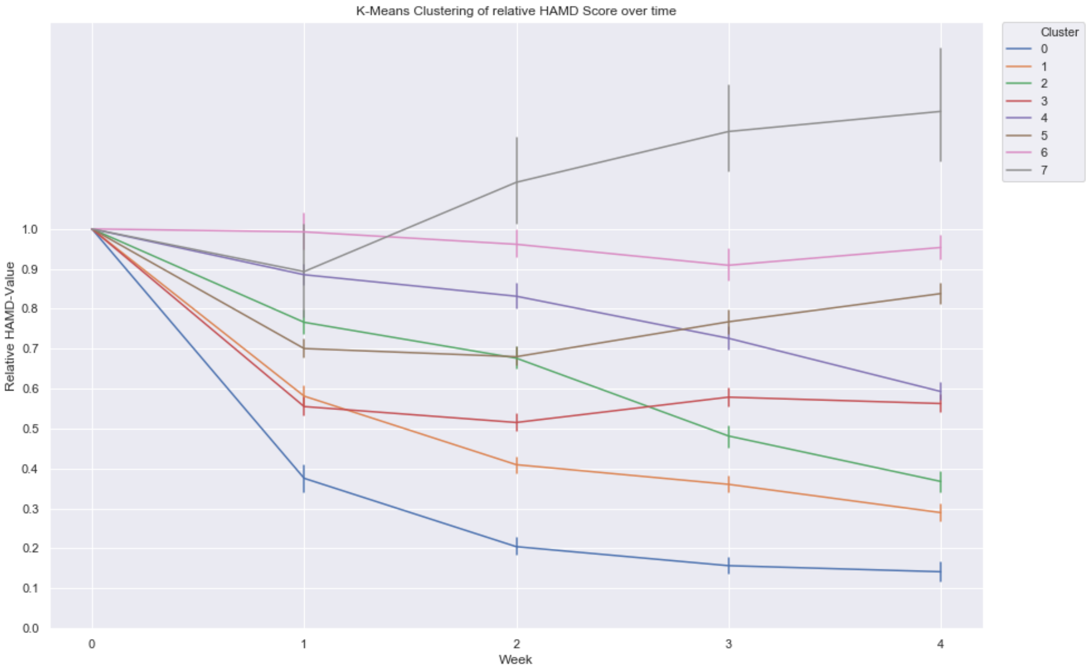

Silhouette Plot for K-Means Clustering (Validation)
Silhouette-Score
−0.2
−0.1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Cluster
0
1
2
3
4
5

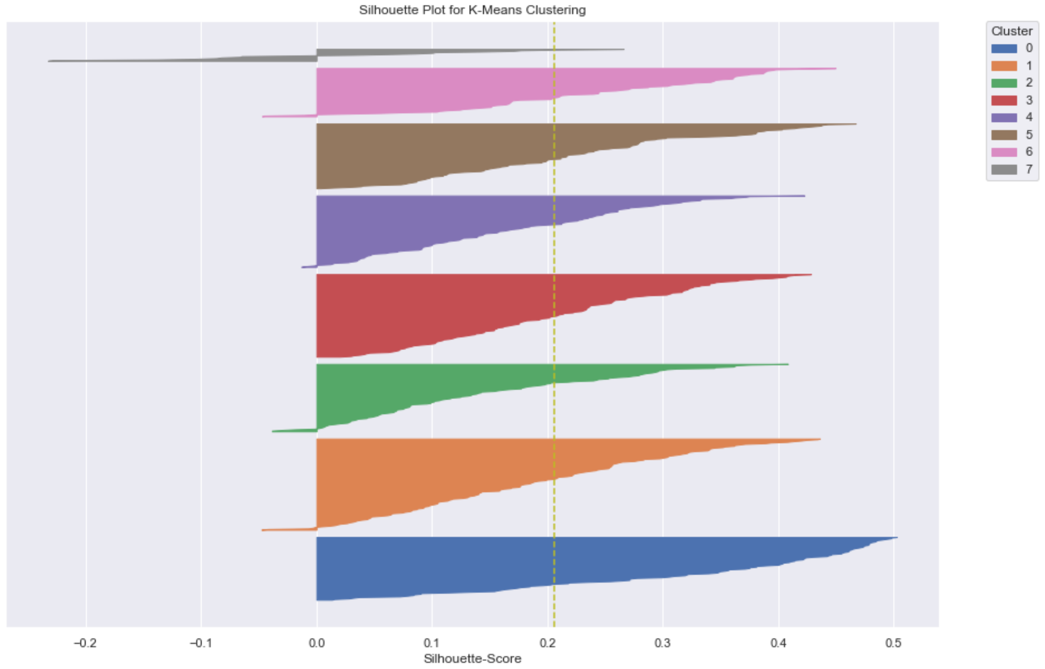

K-Means Clustering of relative HAMD Score over time
Relative HAMD-Value
Week
0
1
2
3
4
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Cluster
0
1
2
3
4
5
6

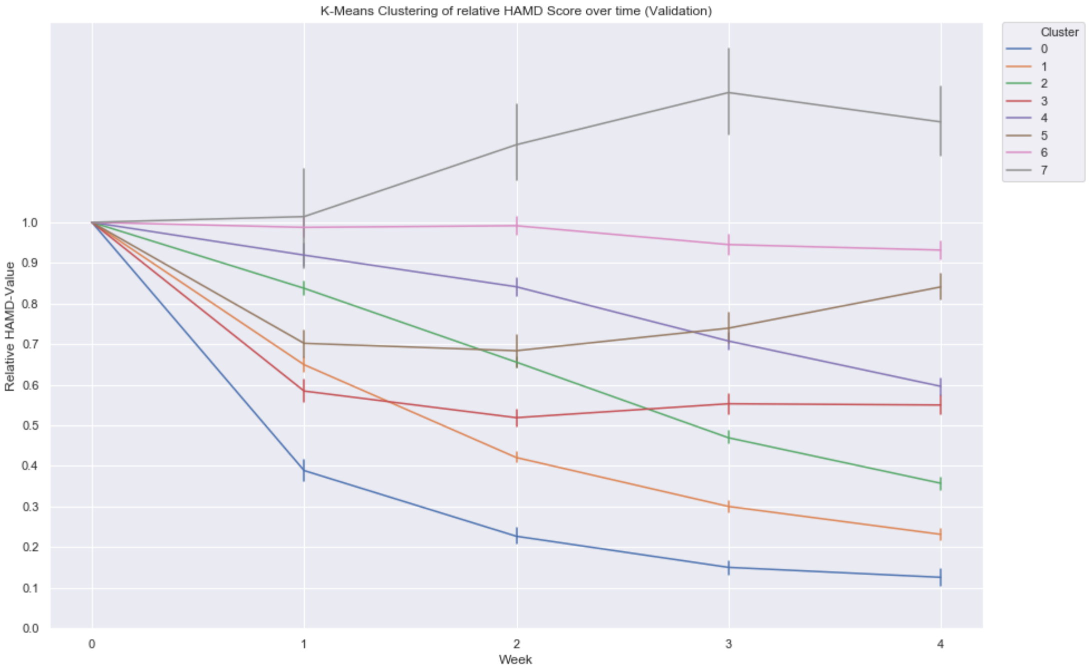

Silhouette Plot for K-Means Clustering
Silhouette-Score
−0.1
0.0
0.1
0.2
0.3
0.4
0.5
Cluster
0
1
2
3
4
5
6

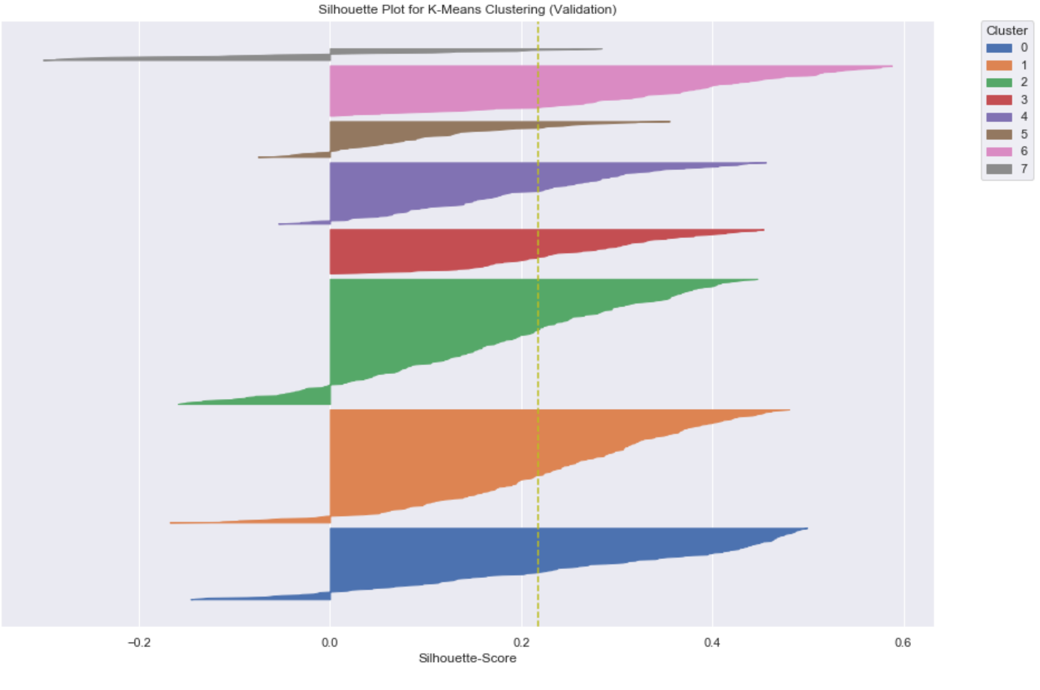

K-Means Clustering of relative HAMD Score over time (Validation)
Relative HAMD-Value
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
1
2
3
4
Week
Cluster
0
1
2
3
4
5
6

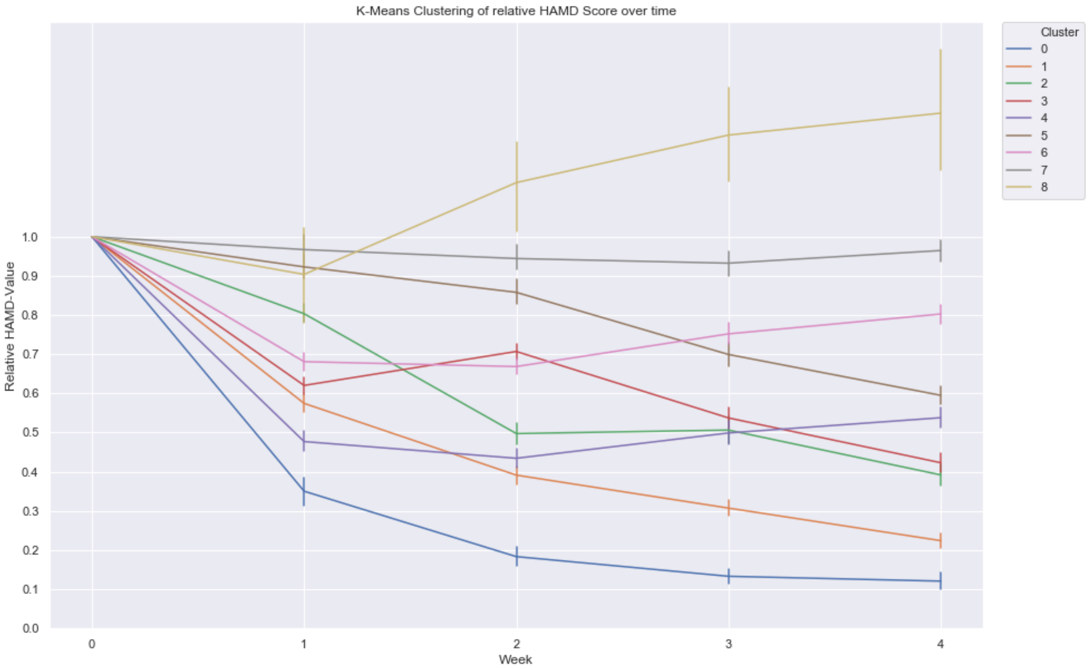

Silhouette Plot for K-Means Clustering (Validation)
−0.2
−0.1
0.0
0.1
0.2
0.3
0.4
0.5
Silhouette-Score
Cluster
0
1
2
3
4
5
6

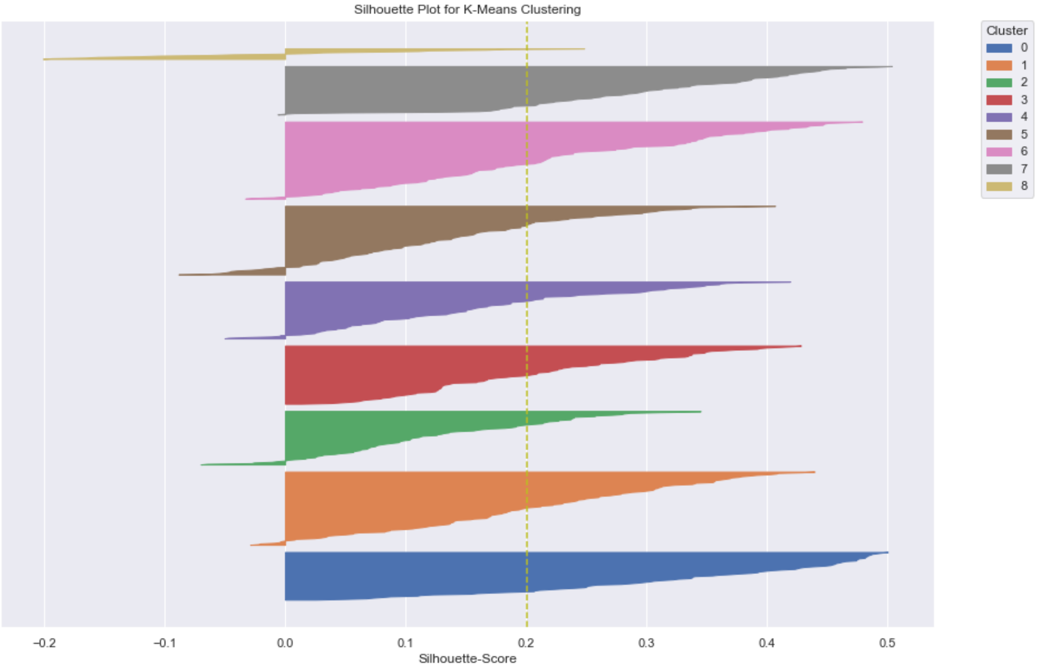

K-Means Clustering of relative HAMD Score over time
Relative HAMD-Value
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
1
2
3
4
Week
Cluster
0
1
2
3
4
5
6
7

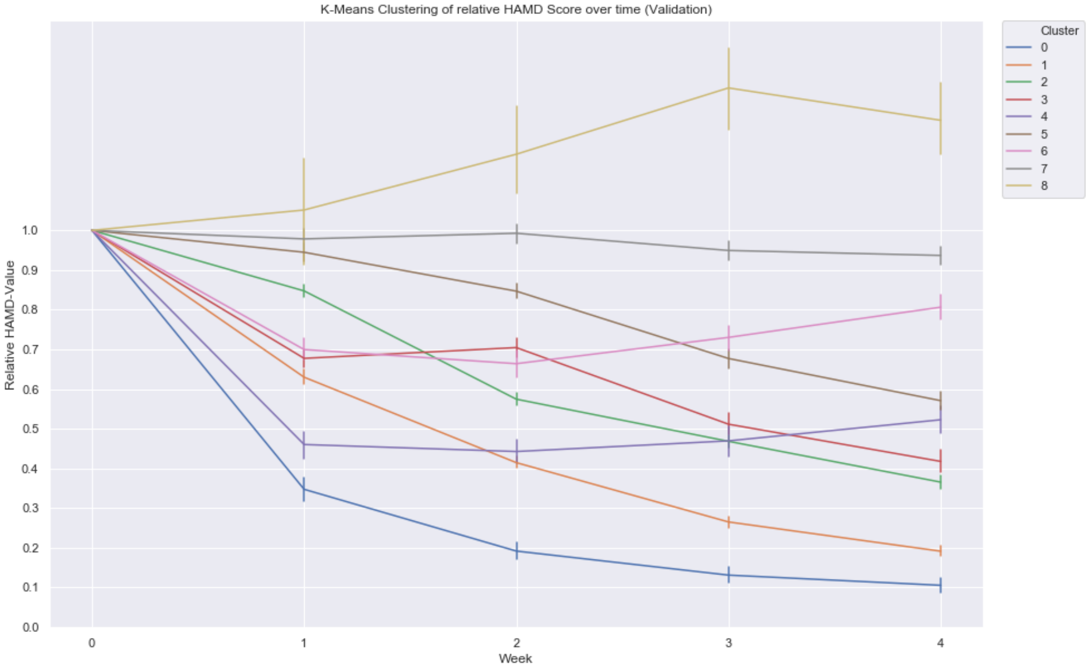

Silhouette Plot for K-Means Clustering
Cluster
0
1
2
3
4
5
6
7
−0.2
−0.1
0.0
0.1
0.2
0.3
0.4
0.5
Silhouette-Score

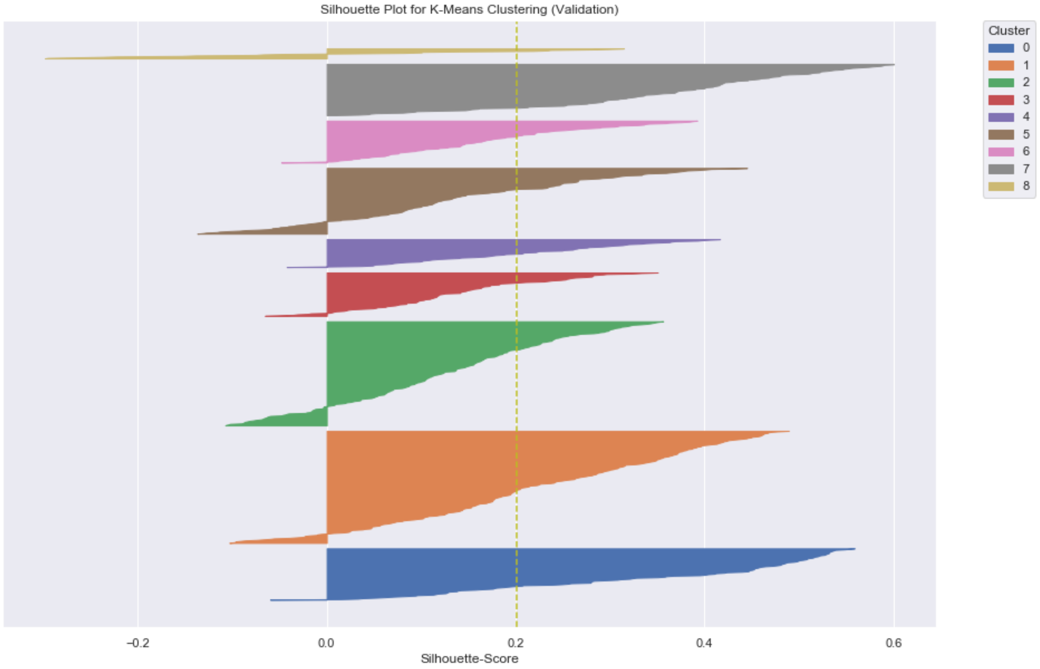

K-Means Clustering of relative HAMD Score over time (Validation)
Relative HAMD-Value
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Week
0
1
2
3
4
Cluster
0
1
2
3
4
5
6
7

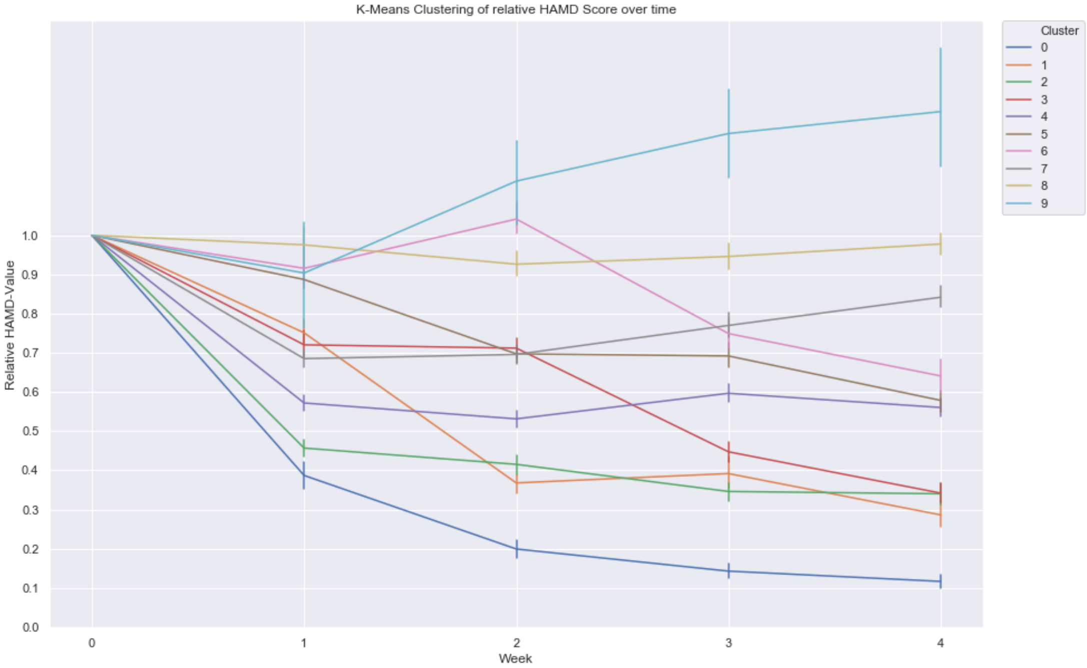

Silhouette Plot for K-Means Clustering (Validation)
Silhouette-Score
−0.2
0.0
0.2
0.4
0.6
Cluster
0
1
2
3
4
5
6
7

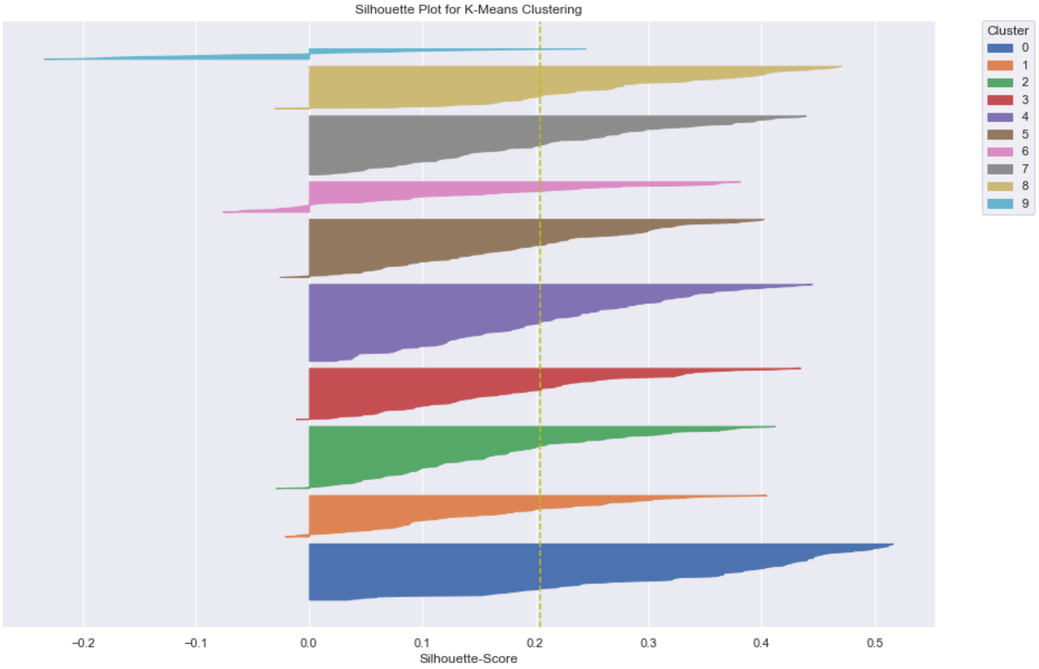

K-Means Clustering of relative HAMD Score over time
Relative HAMD-Value
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
1
2
3
4
Week
Cluster
0
1
2
3
4
5
6
7
8

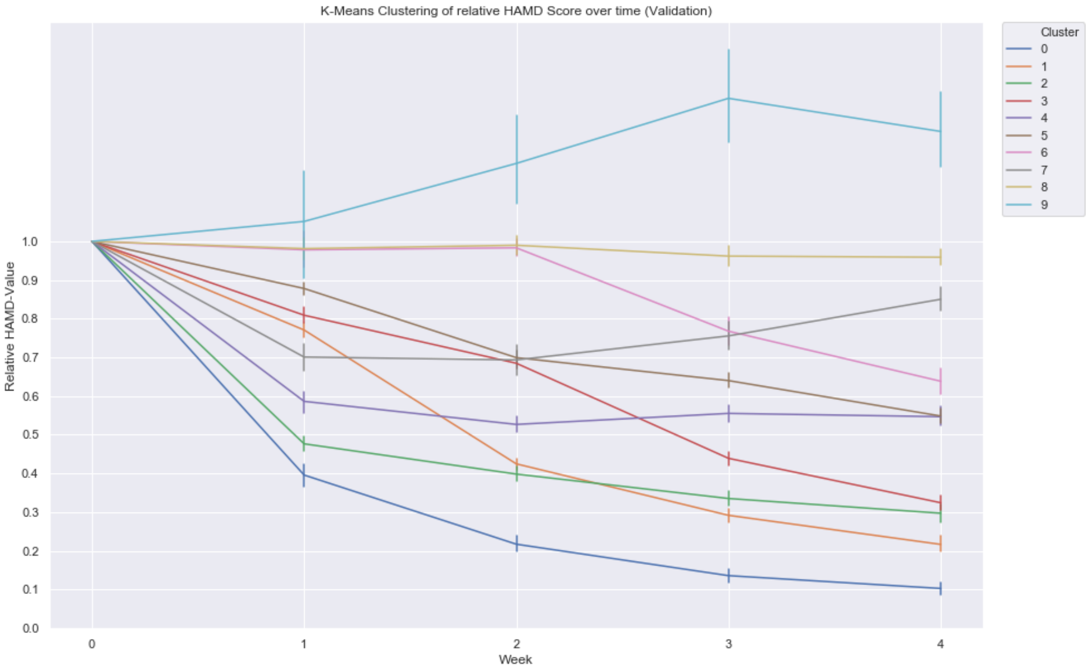

Silhouette Plot for K-Means Clustering
Silhouette-Score
−0.2
−0.1
0.0
0.1
0.2
0.3
0.4
0.5
Cluster
0
1
2
3
4
5
6
7
8

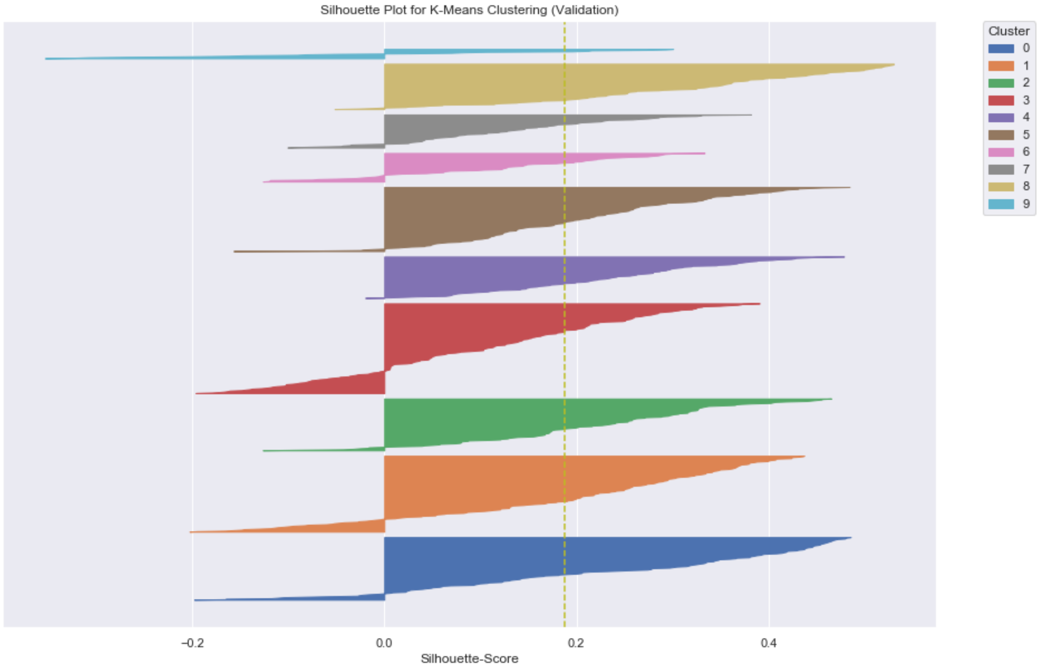

K-Means Clustering of relative HAMD Score over time (Validation)
Relative HAMD-Value
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Week
0
1
2
3
4
Cluster
0
1
2
3
4
5
6
7
8


Silhouette Plot for K-Means Clustering (Validation)
Silhouette-Score
−0.2
0.0
0.2
0.4
0.6
Cluster
0
1
2
3
4
5
6
7
8

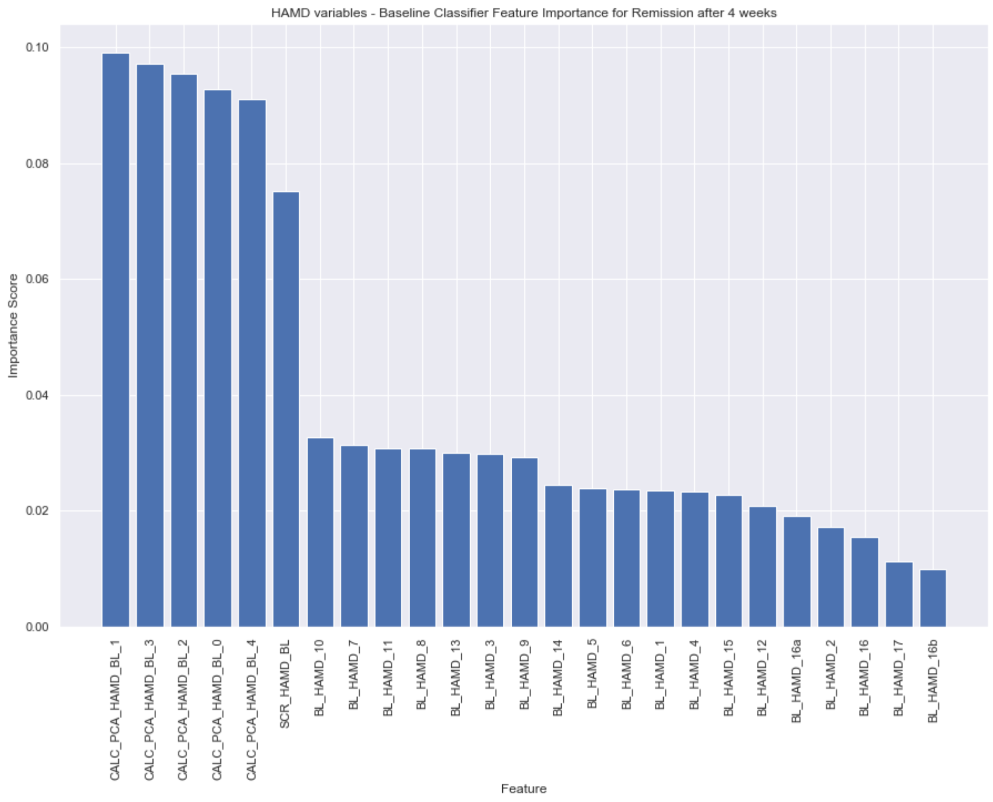

K-Means Clustering of relative HAMD Score over time
Relative HAMD-Value
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
1
2
3
4
Week
Cluster
0
1
2
3
4
5
6
7
8
9

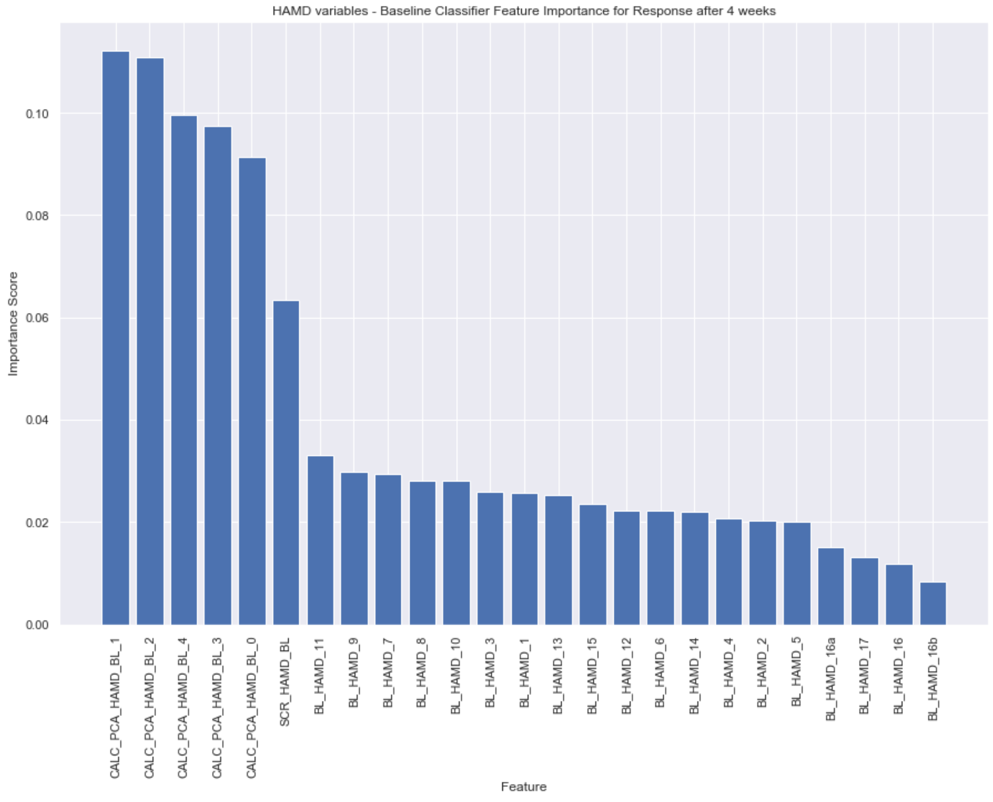

Silhouette Plot for K-Means Clustering
−0.2
−0.1
0.0
0.1
0.2
0.3
0.4
0.5
Silhouette-Score
Cluster
0
1
2
3
4
5
6
7
8
9

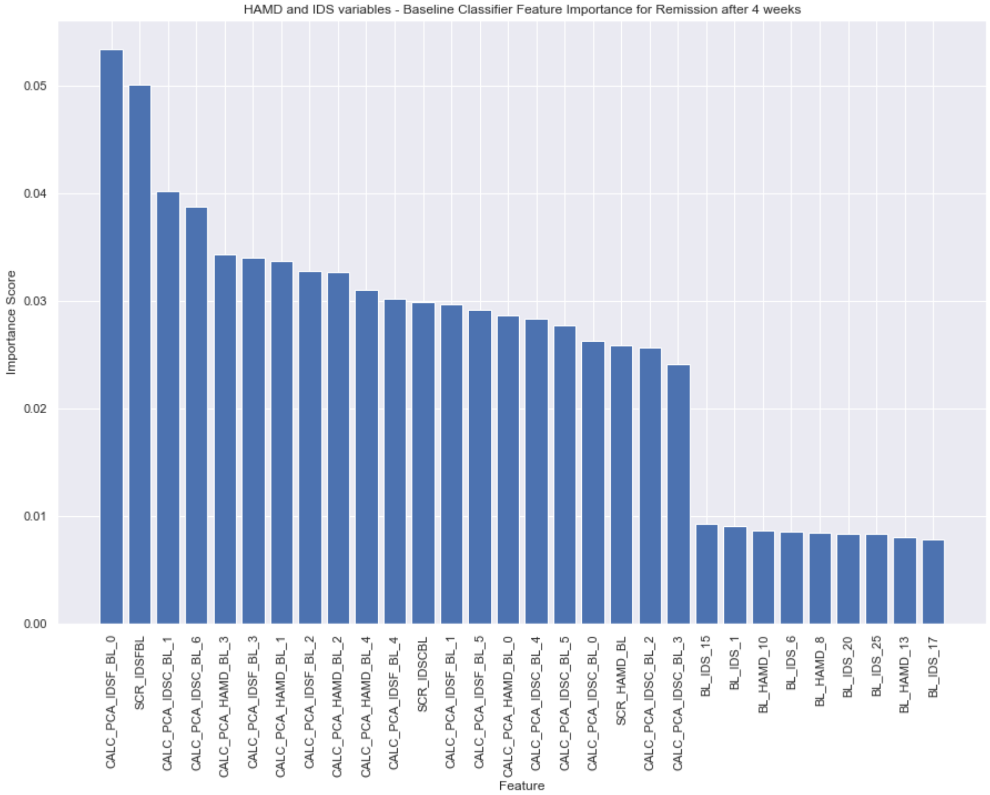

K-Means Clustering of relative HAMD Score over time (Validation)
Relative HAMD-Value
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
1
2
3
4
Week
Cluster
0
1
2
3
4
5
6
7
8
9

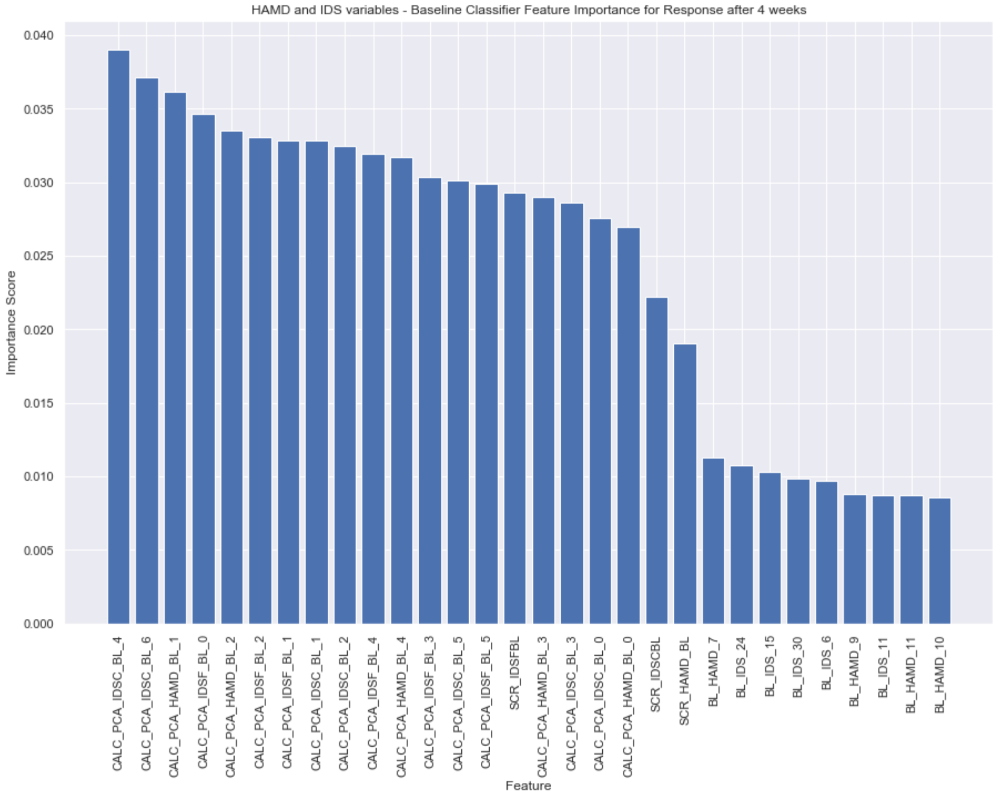

Silhouette Plot for K-Means Clustering (Validation)
Cluster
0
1
2
3
4
5
6
7
8
9
−0.2
0.0
0.2
0.4
Silhouette-Score

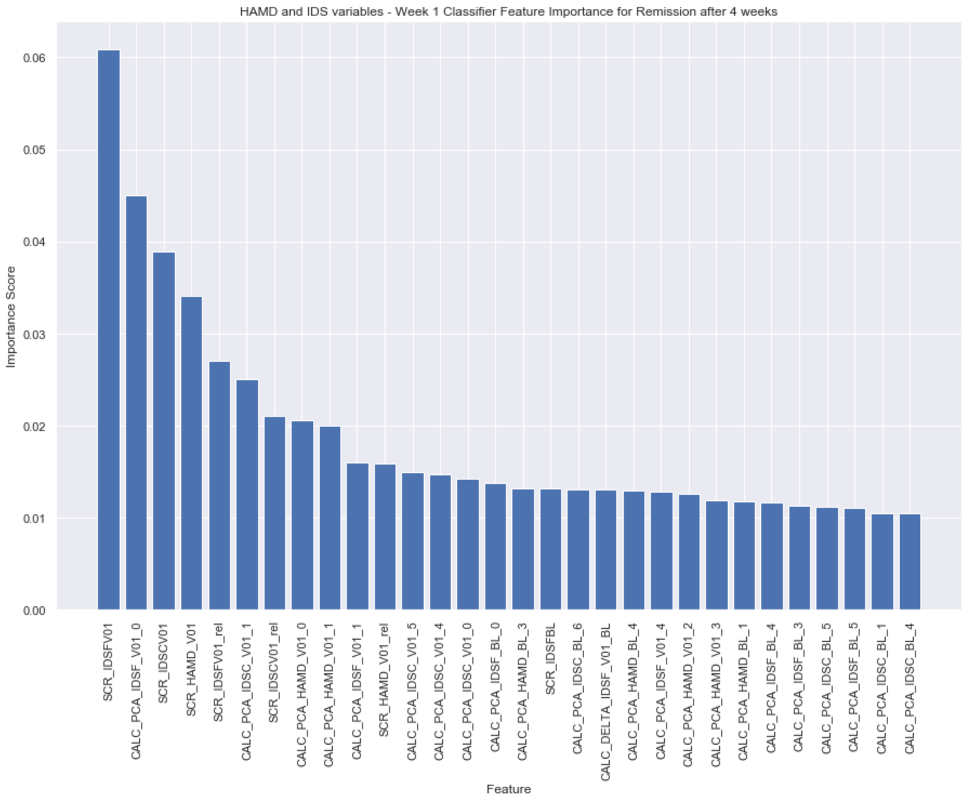
Experiment 2



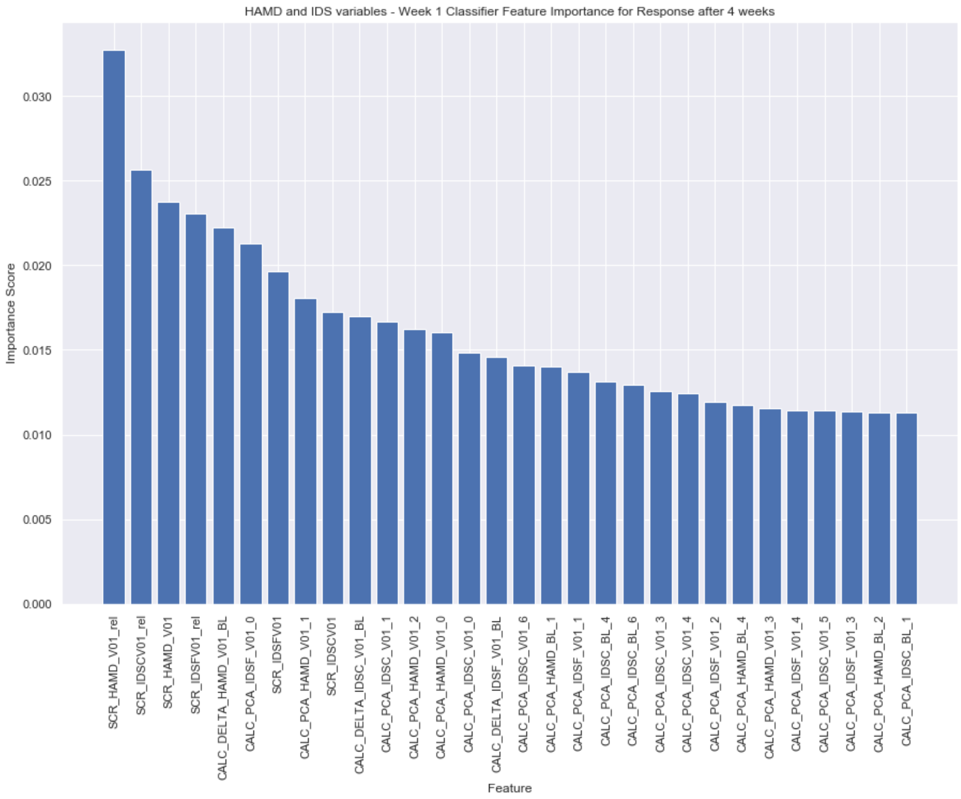



HAMD and IDS variables - Baseline Classifier Feature Importance for Remission after 4 weeks


HAMD and IDS variables - Baseline Classifier Feature Importance for Response after 4 weeks

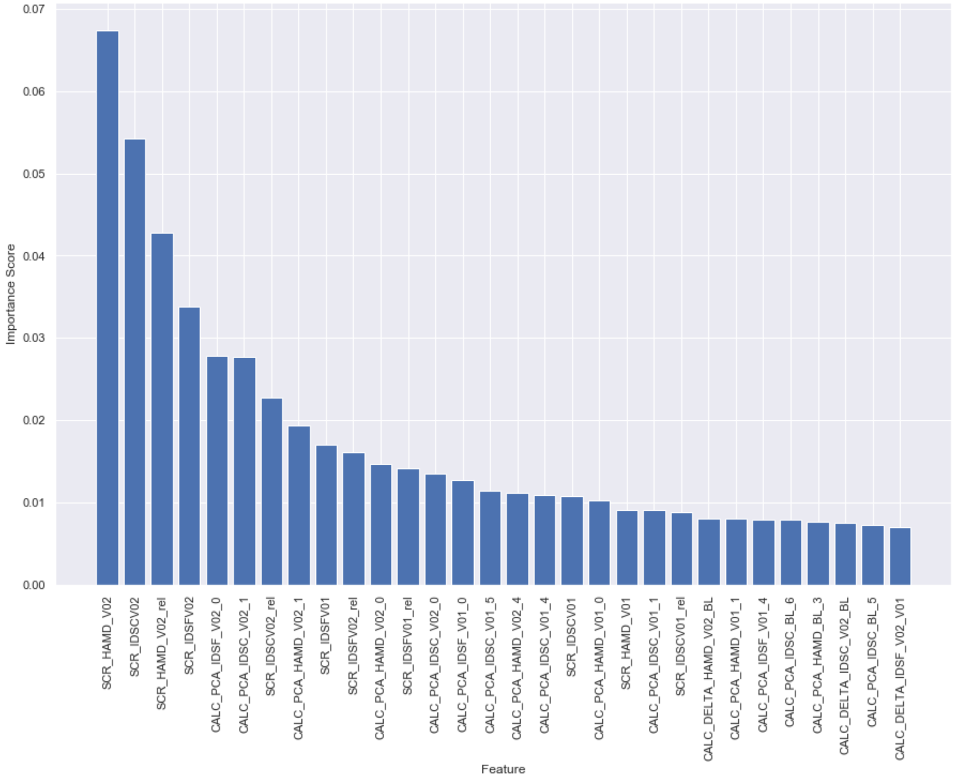


HAMD and IDS variables - Week 1 Classifier Feature Importance for Remission after 4 weeks

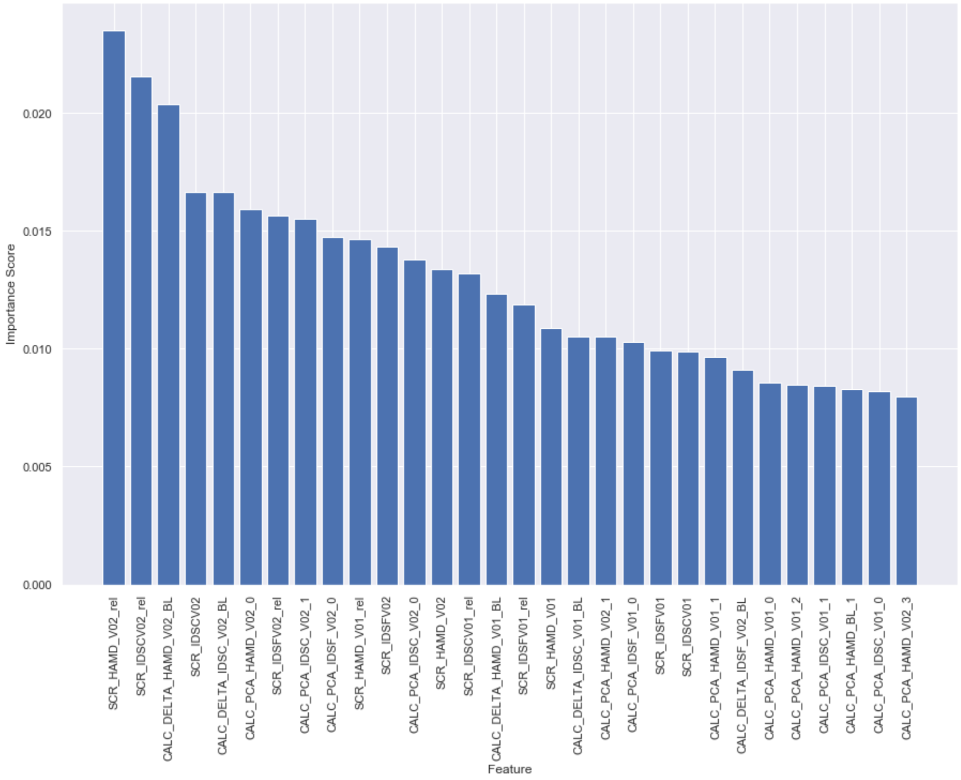


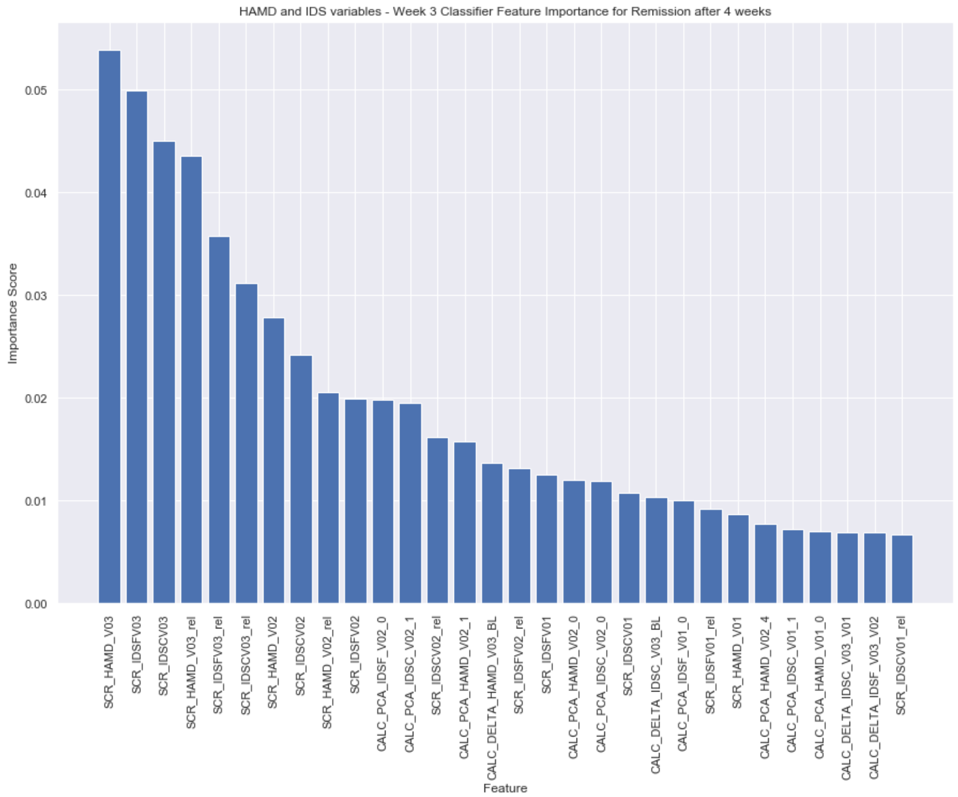
HAMD and IDS variables - Week 1 Classifier Feature Importance for Response after 4 weeks

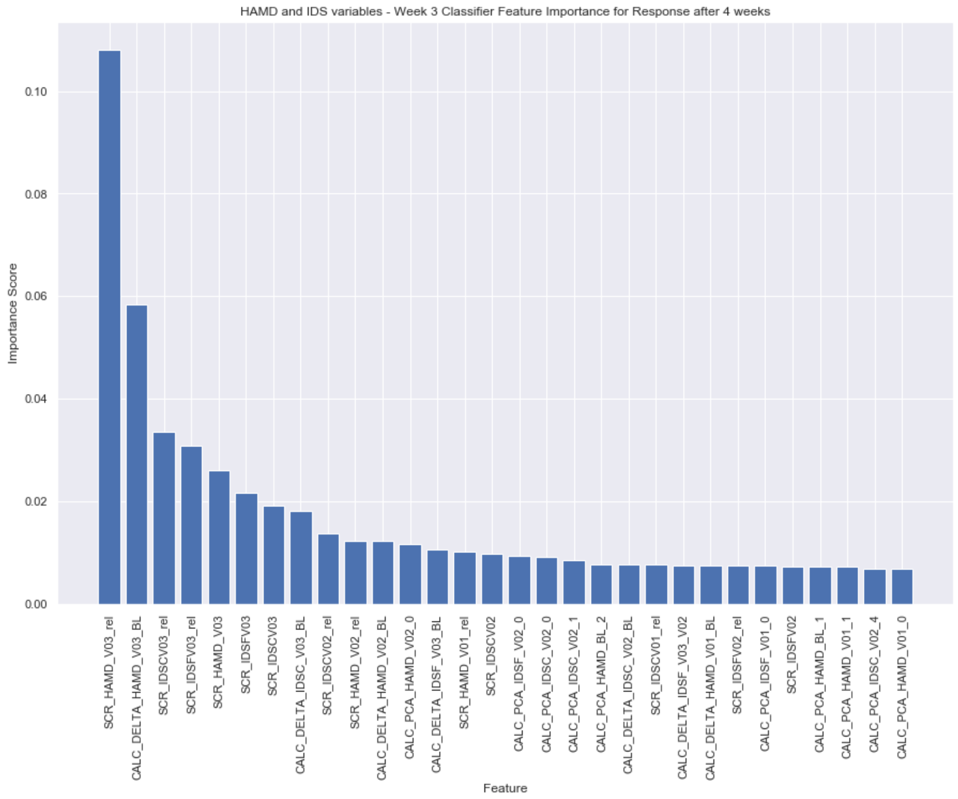
HAMD and IDS variables - Week 2 Classifier Feature Importance for Remission after 4 weeks

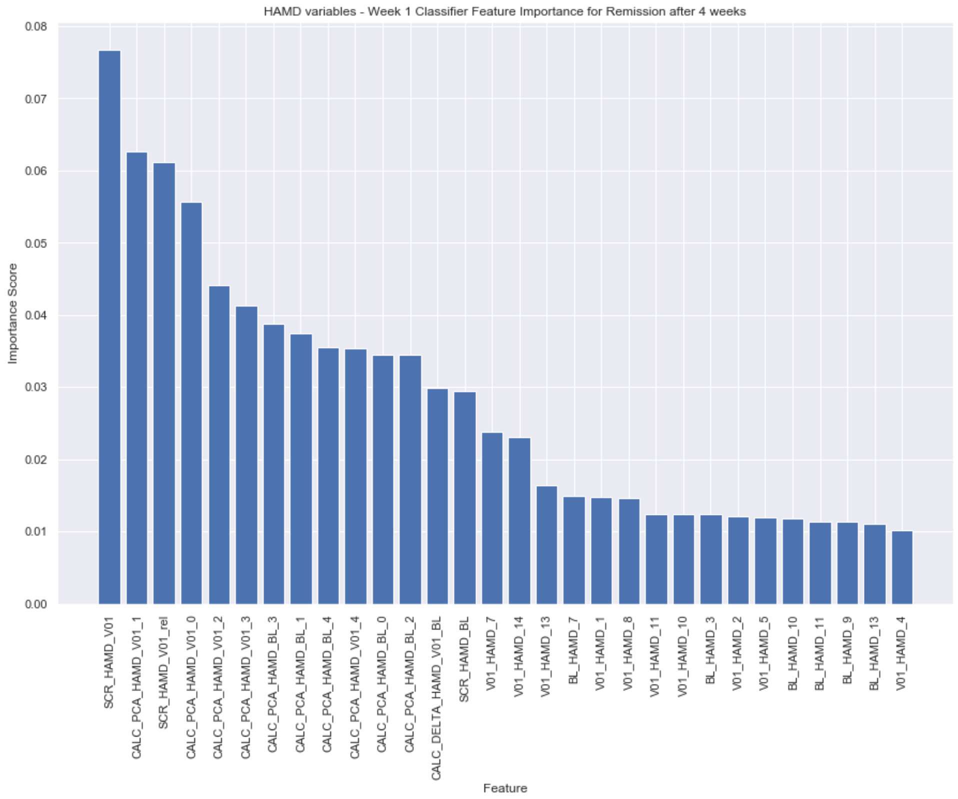


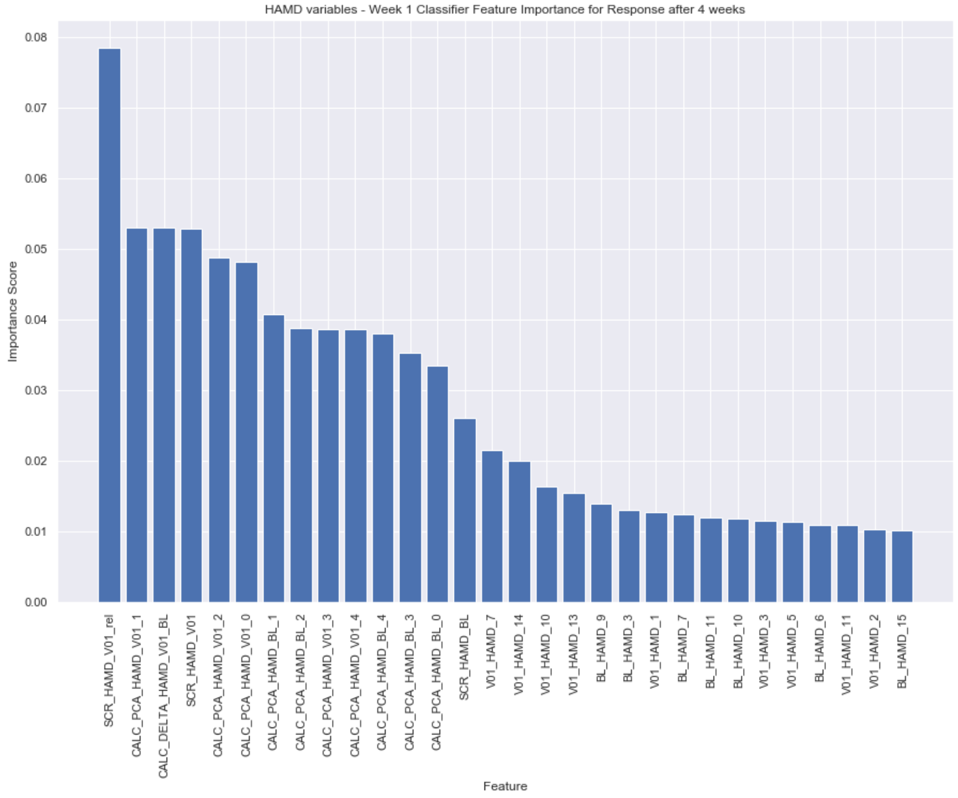
HAMD and IDS variables - Week 2 Classifier Feature Importance for Response after 4 weeks

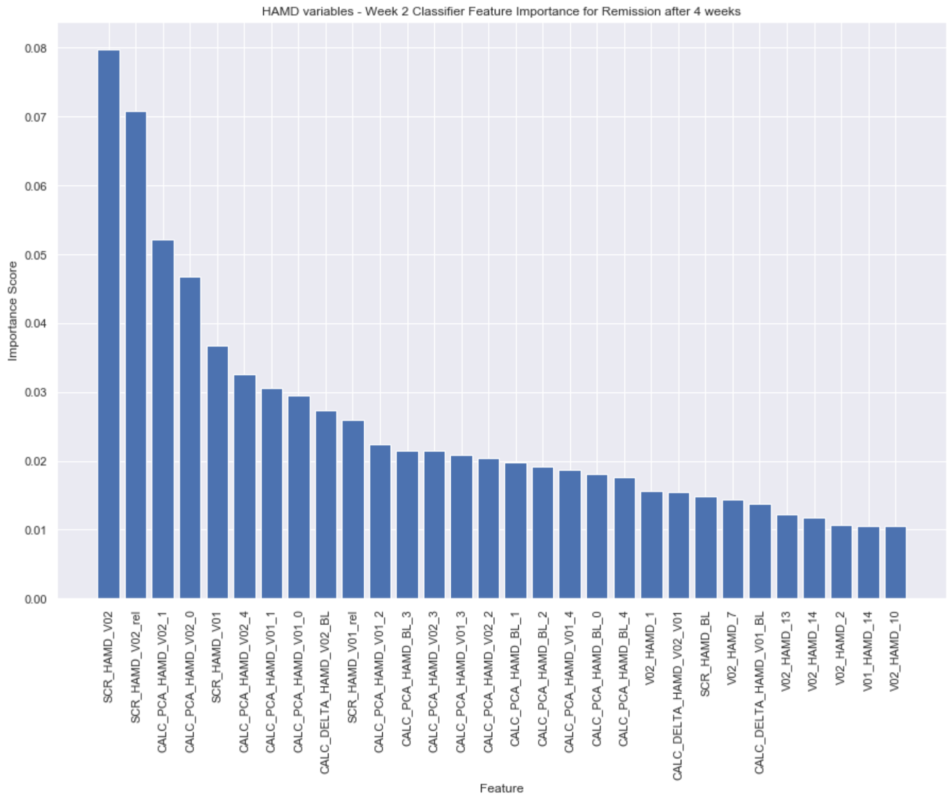



HAMD and IDS variables - Week 3 Classifier Feature Importance for Remission after 4 weeks
Importance Score
0.00
0.01
0.02
0.03
0.04
0.05
Feature
SCR_HAMD_V03
SCR_IDSFV03
SCR_IDSCV03
SCR_HAMD_V03_rel
SCR_IDSFV03_rel
SCR_IDSCV03_rel
SCR_HAMD_V02
SCR_IDSCV02
SCR_HAMD_V02_rel
SCR_IDSFV02
CALC_PCA_IDSF_V02_0
CALC_PCA_IDSC_V02_1
SCR_IDSCV02_rel
CALC_PCA_HAMD_V02_1
CALC_DELTA_HAMD_V03_BL
SCR_IDSFV02_rel
SCR_IDSFV01
CALC_PCA_HAMD_V02_0
CALC_PCA_IDSC_V02_0
SCR_IDSCV01
CALC_DELTA_IDSC_V03_BL
CALC_PCA_IDSF_V01_0
SCR_IDSFV01_rel
SCR_HAMD_V01
CALC_PCA_HAMD_V02_4
CALC_PCA_IDSC_V01_1
CALC_PCA_HAMD_V01_0
CALC_DELTA_IDSC_V03_V01
CALC_DELTA_IDSF_V03_V02
SCR_IDSCV01_rel

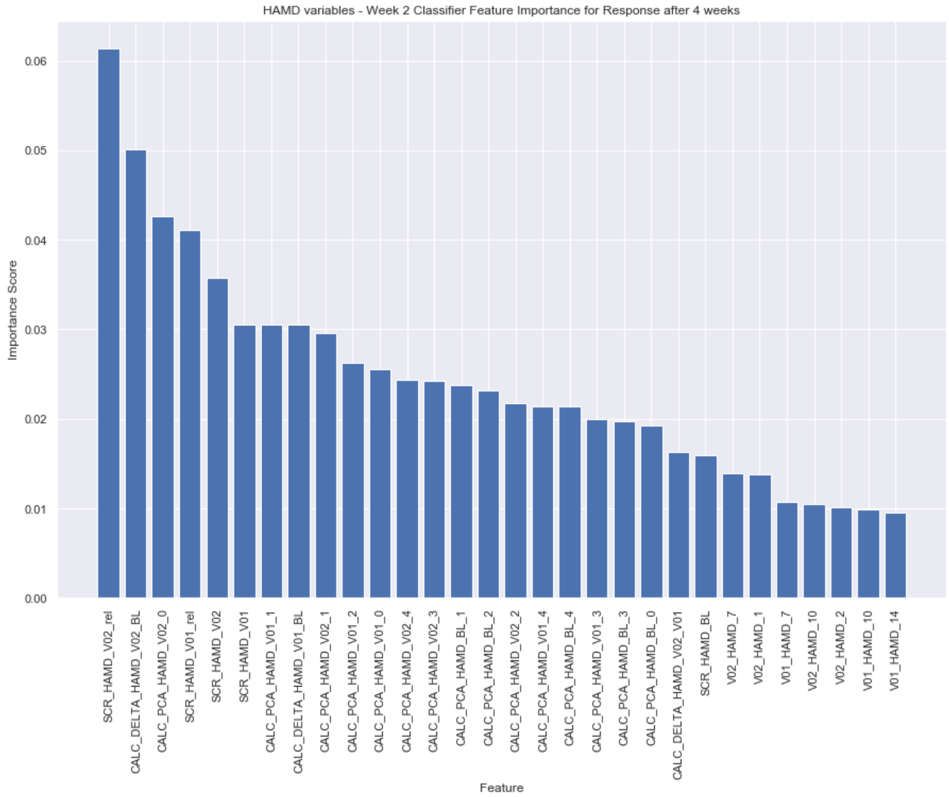

HAMD and IDS variables - Week 3 Classifier Feature Importance for Response after 4 weeks
Importance Score
0.00
0.02
0.04
0.06
0.08
0.10
Feature
SCR_HAMD_V03_rel
CALC_DELTA_HAMD_V03_BL
SCR_IDSCV03_rel
SCR_IDSFV03_rel
SCR_HAMD_V03
SCR_IDSFV03
SCR_IDSCV03
CALC_DELTA_IDSC_V03_BL
SCR_IDSCV02_rel
SCR_HAMD_V02_rel
CALC_DELTA_HAMD_V02_BL
CALC_PCA_HAMD_V02_0
CALC_DELTA_IDSF_V03_BL
SCR_HAMD_V01_rel
SCR_IDSCV02
CALC_PCA_IDSF_V02_0
CALC_PCA_IDSC_V02_0
CALC_PCA_IDSC_V02_1
CALC_PCA_HAMD_BL_2
CALC_DELTA_IDSC_V02_BL
SCR_IDSCV01_rel
CALC_DELTA_IDSF_V03_V02
CALC_DELTA_HAMD_V01_BL
SCR_IDSFV02_rel
CALC_PCA_IDSF_V01_0
SCR_IDSFV02
CALC_PCA_HAMD_BL_1
CALC_PCA_HAMD_V01_1
CALC_PCA_IDSC_V02_4
CALC_PCA_HAMD_V01_0


HAMD variables - Week 1 Classifier Feature Importance for Remission after 4 weeks
Importance Score
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
SCR_HAMD_V01
CALC_PCA_HAMD_V01_1
SCR_HAMD_V01_rel
CALC_PCA_HAMD_V01_0
CALC_PCA_HAMD_V01_2
CALC_PCA_HAMD_V01_3
CALC_PCA_HAMD_BL_3
CALC_PCA_HAMD_BL_1
CALC_PCA_HAMD_BL_4
CALC_PCA_HAMD_V01_4
CALC_PCA_HAMD_BL_0
CALC_PCA_HAMD_BL_2
CALC_DELTA_HAMD_V01_BL
SCR_HAMD_BL
V01_HAMD_7
V01_HAMD_14
V01_HAMD_13
BL_HAMD_7
V01_HAMD_1
V01_HAMD_8
V01_HAMD_11
V01_HAMD_10
BL_HAMD_3
V01_HAMD_2
V01_HAMD_5
BL_HAMD_10
BL_HAMD_11
BL_HAMD_9
BL_HAMD_13
V01_HAMD_4
Feature


HAMD variables - Week 1 Classifier Feature Importance for Response after 4 weeks
Importance Score
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
SCR_HAMD_V01_rel
CALC_PCA_HAMD_V01_1
CALC_DELTA_HAMD_V01_BL
SCR_HAMD_V01
CALC_PCA_HAMD_V01_2
CALC_PCA_HAMD_V01_0
CALC_PCA_HAMD_BL_1
CALC_PCA_HAMD_BL_2
CALC_PCA_HAMD_V01_3
CALC_PCA_HAMD_V01_4
CALC_PCA_HAMD_BL_4
CALC_PCA_HAMD_BL_3
CALC_PCA_HAMD_BL_0
SCR_HAMD_BL
V01_HAMD_7
V01_HAMD_14
V01_HAMD_10
V01_HAMD_13
BL_HAMD_9
BL_HAMD_3
V01_HAMD_1
BL_HAMD_7
BL_HAMD_11
BL_HAMD_10
V01_HAMD_3
V01_HAMD_5
BL_HAMD_6
V01_HAMD_11
V01_HAMD_2
BL_HAMD_15
Feature

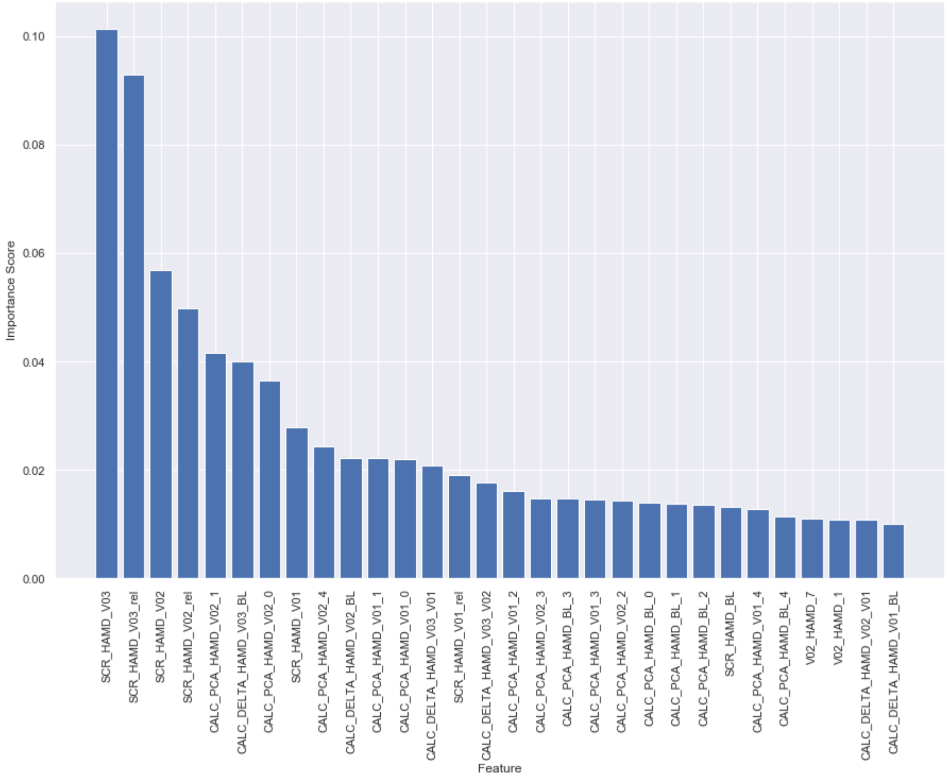


HAMD variables - Week 2 Classifier Feature Importance for Remission after 4 weeks

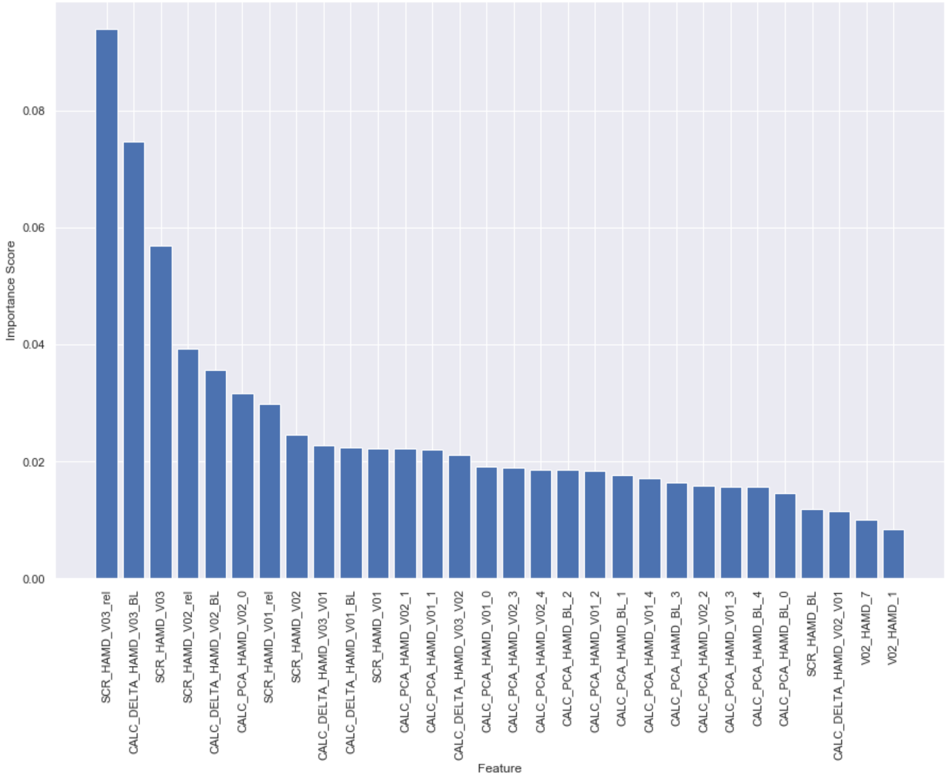


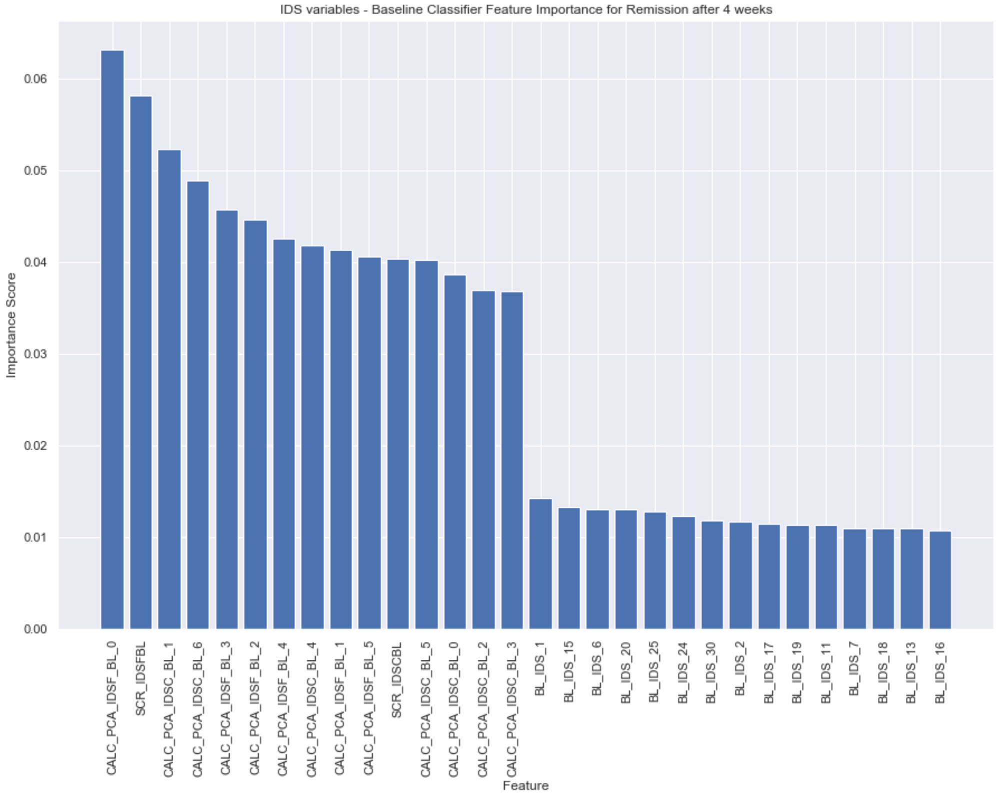
HAMD variables - Week 2 Classifier Feature Importance for Response after 4 weeks

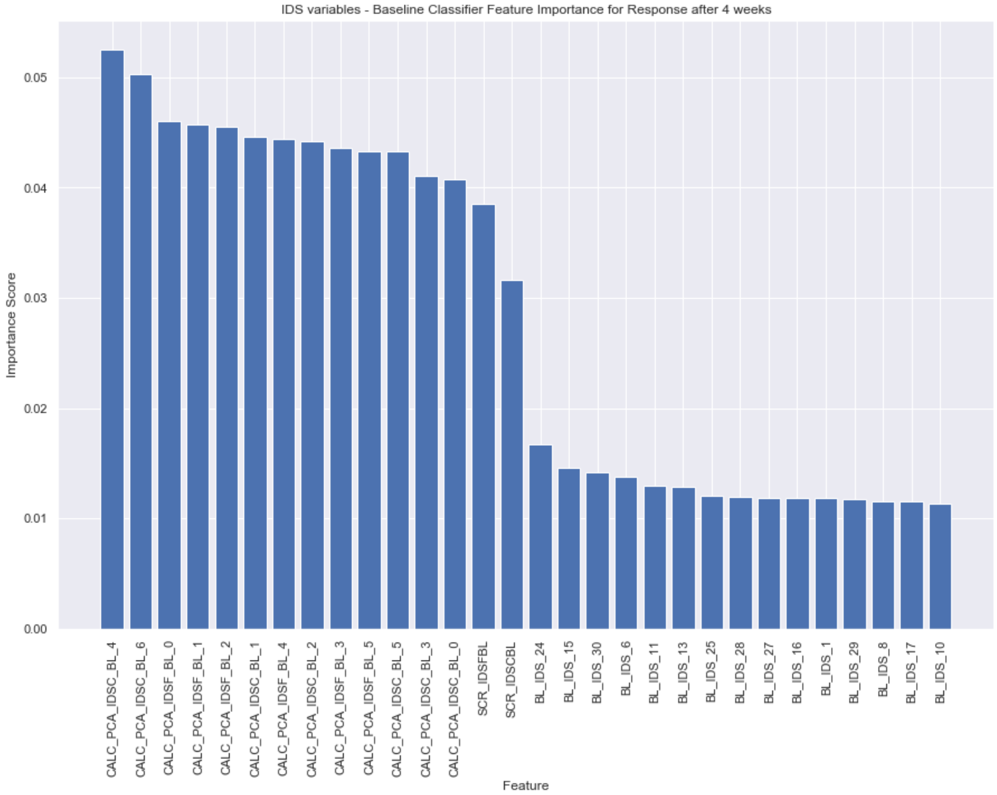
HAMD variables - Week 3 Classifier Feature Importance for Remission after 4 weeks

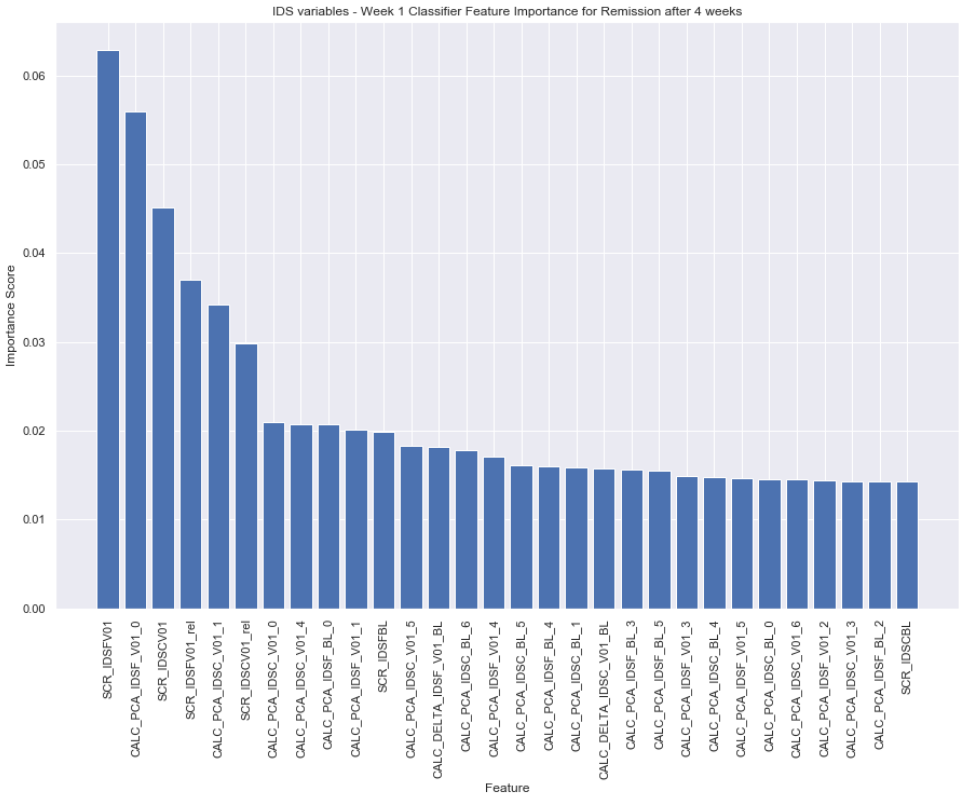


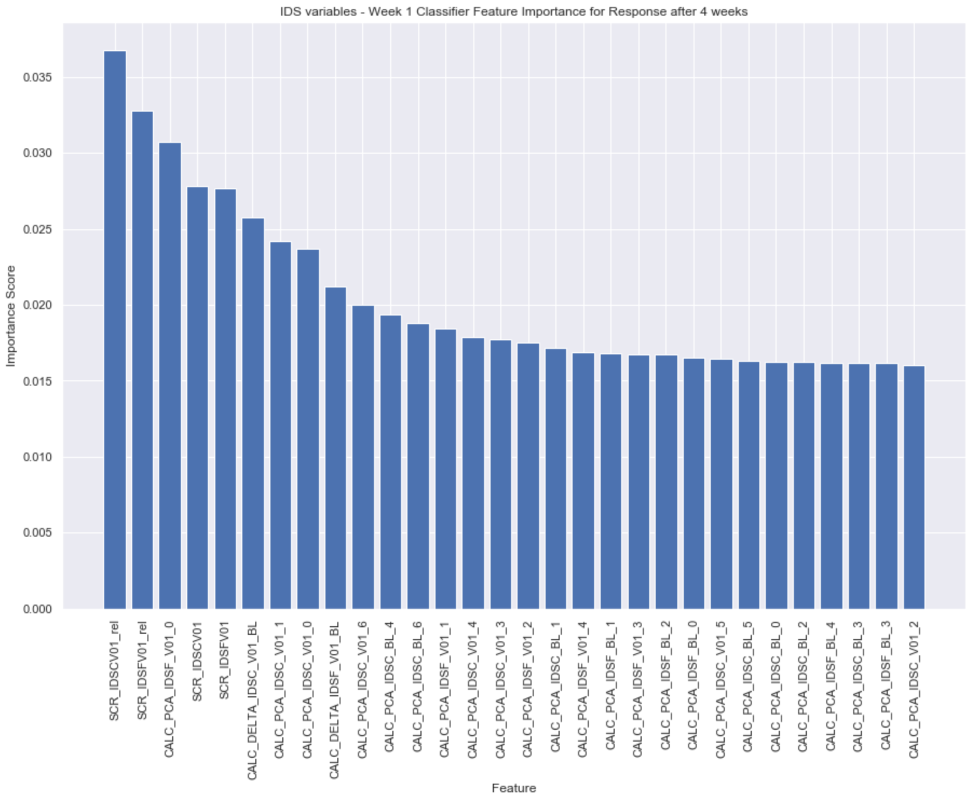
HAMD variables - Week 3 Classifier Feature Importance for Response after 4 weeks

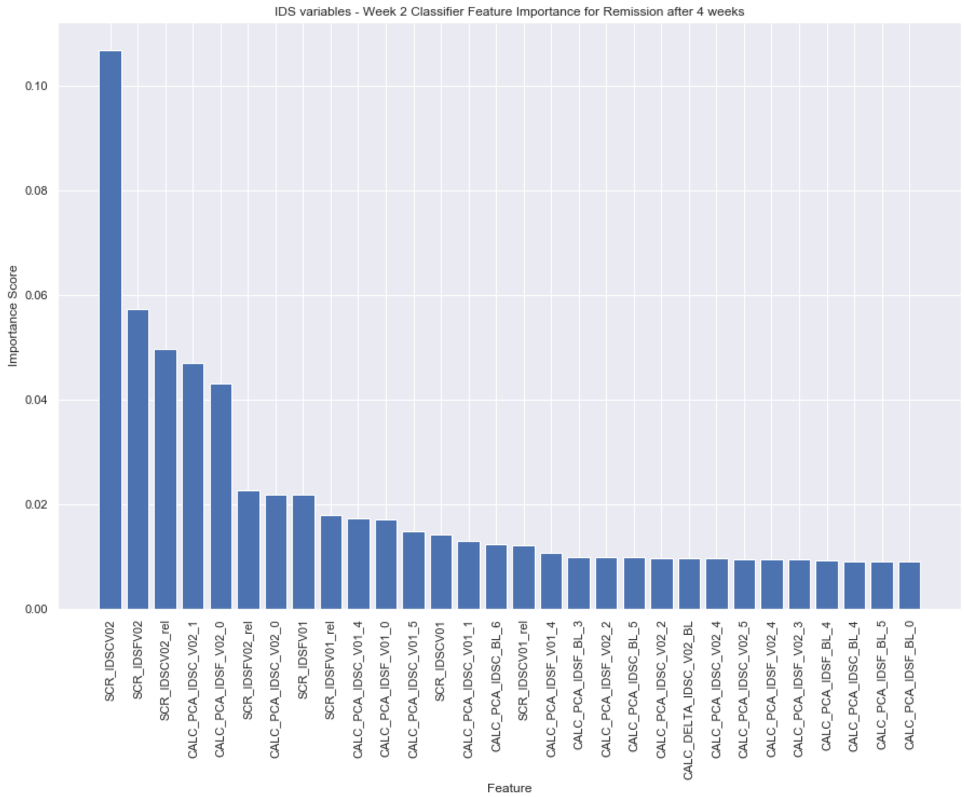



IDS variables - Baseline Classifier Feature Importance for Remission after 4 weeks

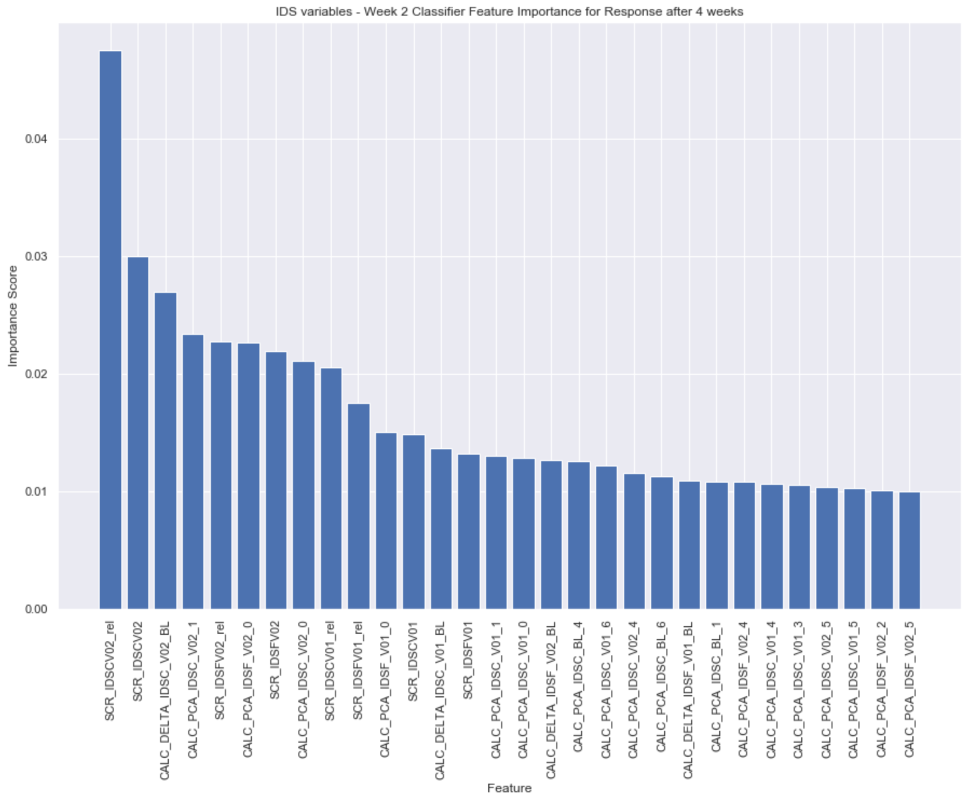

IDS variables - Baseline Classifier Feature Importance for Response after 4 weeks


IDS variables - Week 1 Classifier Feature Importance for Remission after 4 weeks
Importance Score
0.06
0.05
0.04
0.03
0.02
0.01
0.00
SCR_IDSFV01
CALC_PCA_IDSF_V01_0
SCR_IDSCV01
SCR_IDSFV01_rel
CALC_PCA_IDSC_V01_1
SCR_IDSCV01_rel
CALC_PCA_IDSC_V01_0
CALC_PCA_IDSC_V01_4
CALC_PCA_IDSF_BL_0
CALC_PCA_IDSF_V01_1
SCR_IDSFBL
CALC_PCA_IDSC_V01_5
CALC_DELTA_IDSF_V01_BL
CALC_PCA_IDSC_BL_6
CALC_PCA_IDSF_V01_4
CALC_PCA_IDSC_BL_5
CALC_PCA_IDSF_BL_4
CALC_PCA_IDSC_BL_1
CALC_DELTA_IDSC_V01_BL
CALC_PCA_IDSF_BL_3
CALC_PCA_IDSF_BL_5
CALC_PCA_IDSF_V01_3
CALC_PCA_IDSC_BL_4
CALC_PCA_IDSF_V01_5
CALC_PCA_IDSC_BL_0
CALC_PCA_IDSC_V01_6
CALC_PCA_IDSF_V01_2
CALC_PCA_IDSC_V01_3
CALC_PCA_IDSF_BL_2
SCR_IDSCBL
Feature

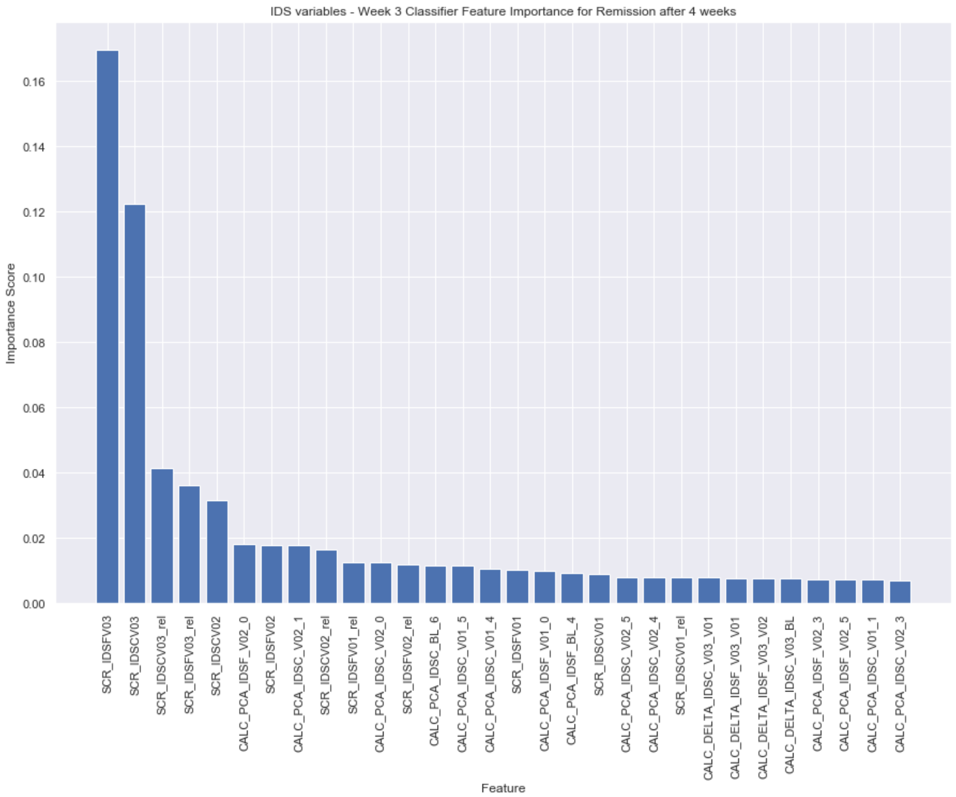

IDS variables - Week 1 Classifier Feature Importance for Response after 4 weeks
Importance Score
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
SCR_IDSCV01_rel
SCR_IDSFV01_rel
CALC_PCA_IDSF_V01_0
SCR_IDSCV01
SCR_IDSFV01
CALC_DELTA_IDSC_V01_BL
CALC_PCA_IDSC_V01_1
CALC_PCA_IDSC_V01_0
CALC_DELTA_IDSF_V01_BL
CALC_PCA_IDSC_V01_6
CALC_PCA_IDSC_BL_4
CALC_PCA_IDSC_BL_6
CALC_PCA_IDSF_V01_1
CALC_PCA_IDSC_V01_4
CALC_PCA_IDSC_V01_3
CALC_PCA_IDSF_V01_2
CALC_PCA_IDSC_BL_1
CALC_PCA_IDSF_V01_4
CALC_PCA_IDSF_BL_1
CALC_PCA_IDSF_V01_3
CALC_PCA_IDSF_BL_2
CALC_PCA_IDSF_BL_0
CALC_PCA_IDSC_V01_5
CALC_PCA_IDSC_BL_5
CALC_PCA_IDSC_BL_0
CALC_PCA_IDSC_BL_2
CALC_PCA_IDSF_BL_4
CALC_PCA_IDSC_BL_3
CALC_PCA_IDSF_BL_3
CALC_PCA_IDSC_V01_2
Feature

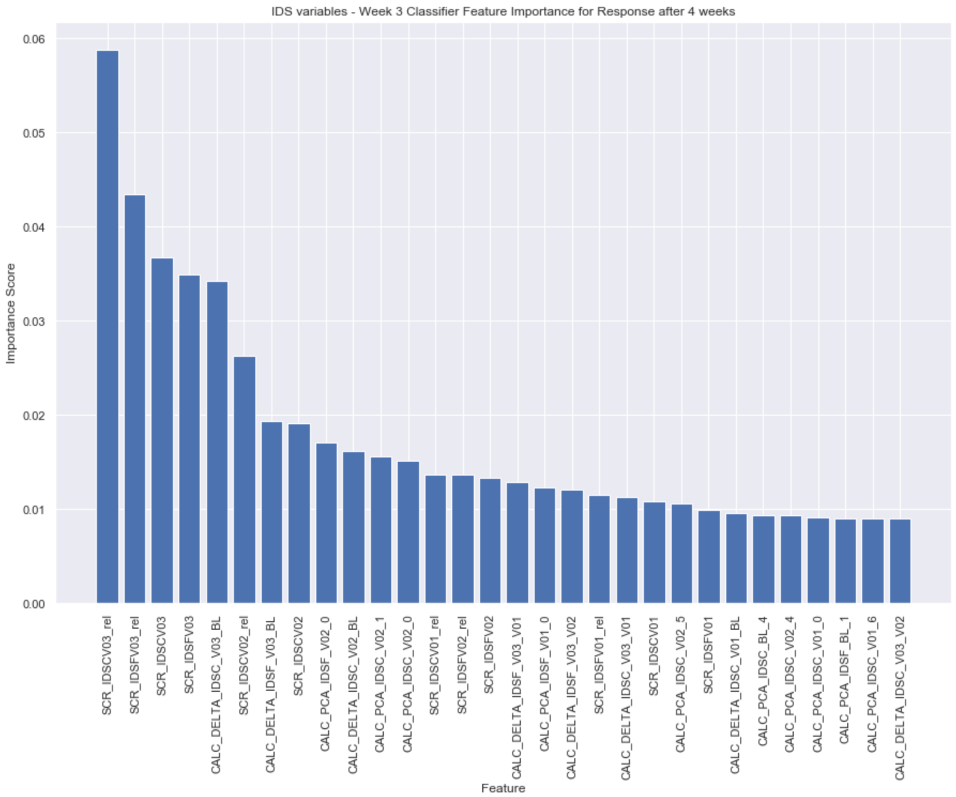


IDS variables - Week 2 Classifier Feature Importance for Remission after 4 weeks

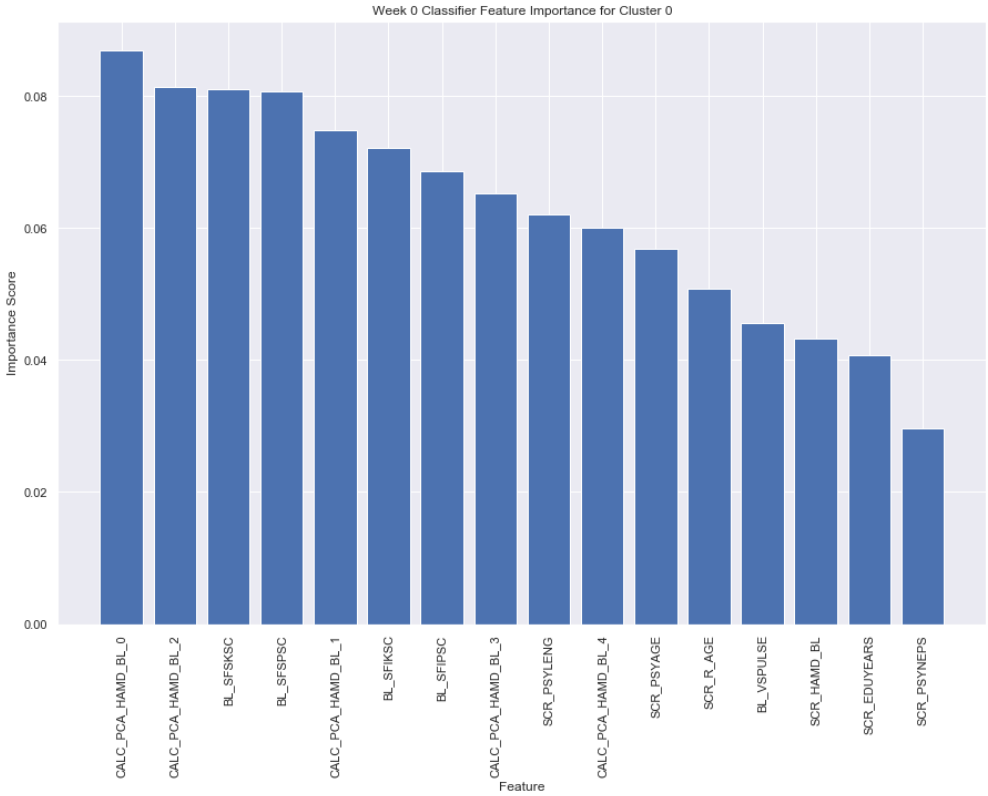


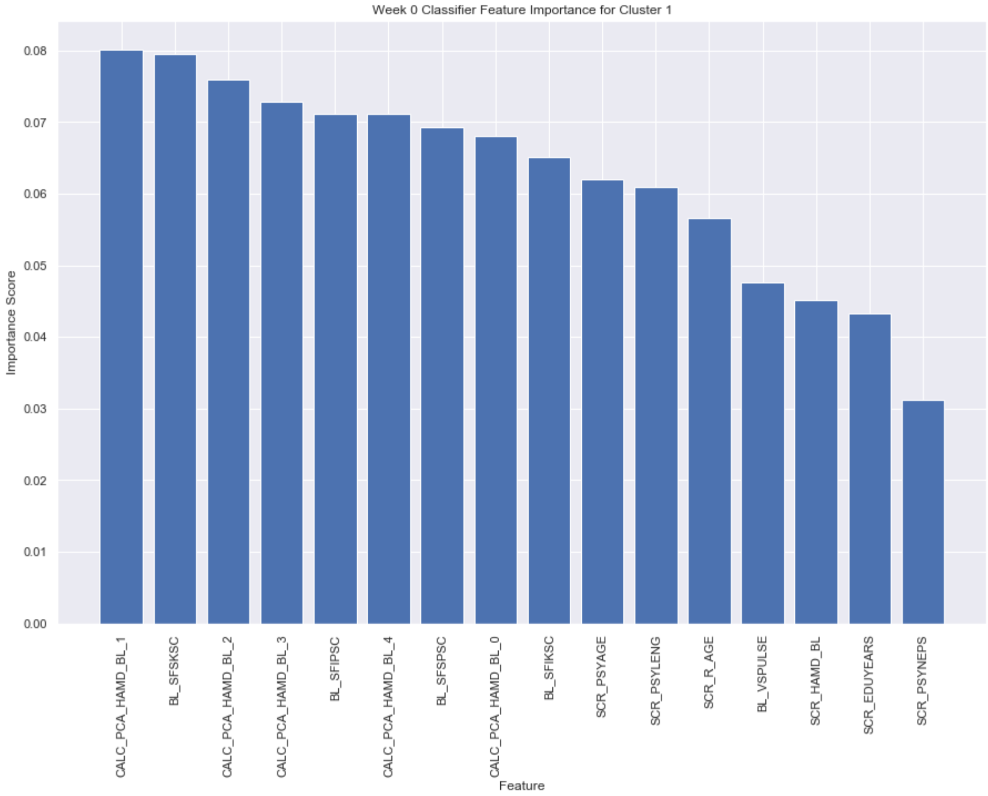
IDS variables - Week 2 Classifier Feature Importance for Response after 4 weeks

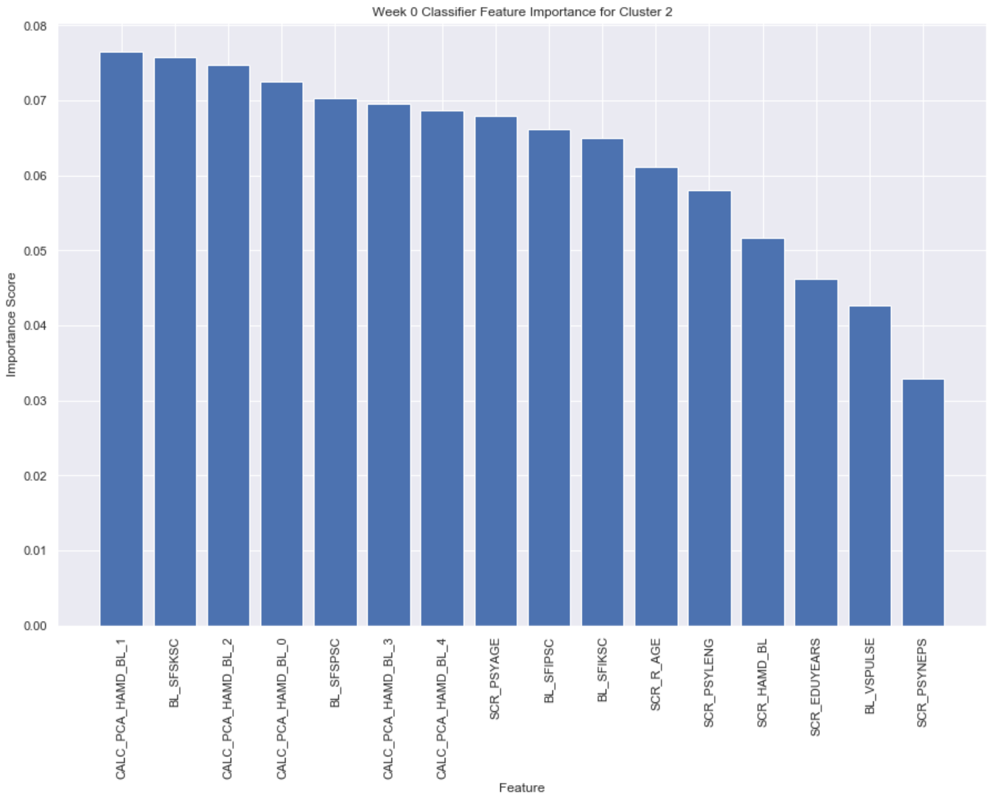

IDS variables - Week 3 Classifier Feature Importance for Remission after 4 weeks
Importance Score
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16
Feature
SCR_IDSFV03
SCR_IDSCV03
SCR_IDSCV03_rel
SCR_IDSFV03_rel
SCR_IDSCV02
CALC_PCA_IDSF_V02_0
SCR_IDSFV02
CALC_PCA_IDSC_V02_1
SCR_IDSCV02_rel
SCR_IDSFV01_rel
CALC_PCA_IDSC_V02_0
SCR_IDSFV02_rel
CALC_PCA_IDSC_BL_6
CALC_PCA_IDSC_V01_5
CALC_PCA_IDSC_V01_4
SCR_IDSFV01
CALC_PCA_IDSF_V01_0
CALC_PCA_IDSF_BL_4
SCR_IDSCV01
CALC_PCA_IDSC_V02_5
CALC_PCA_IDSC_V02_4
SCR_IDSCV01_rel
CALC_DELTA_IDSC_V03_V01
CALC_DELTA_IDSF_V03_V01
CALC_DELTA_IDSF_V03_V02
CALC_DELTA_IDSC_V03_BL
CALC_PCA_IDSF_V02_3
CALC_PCA_IDSF_V02_5
CALC_PCA_IDSC_V01_1
CALC_PCA_IDSC_V02_3

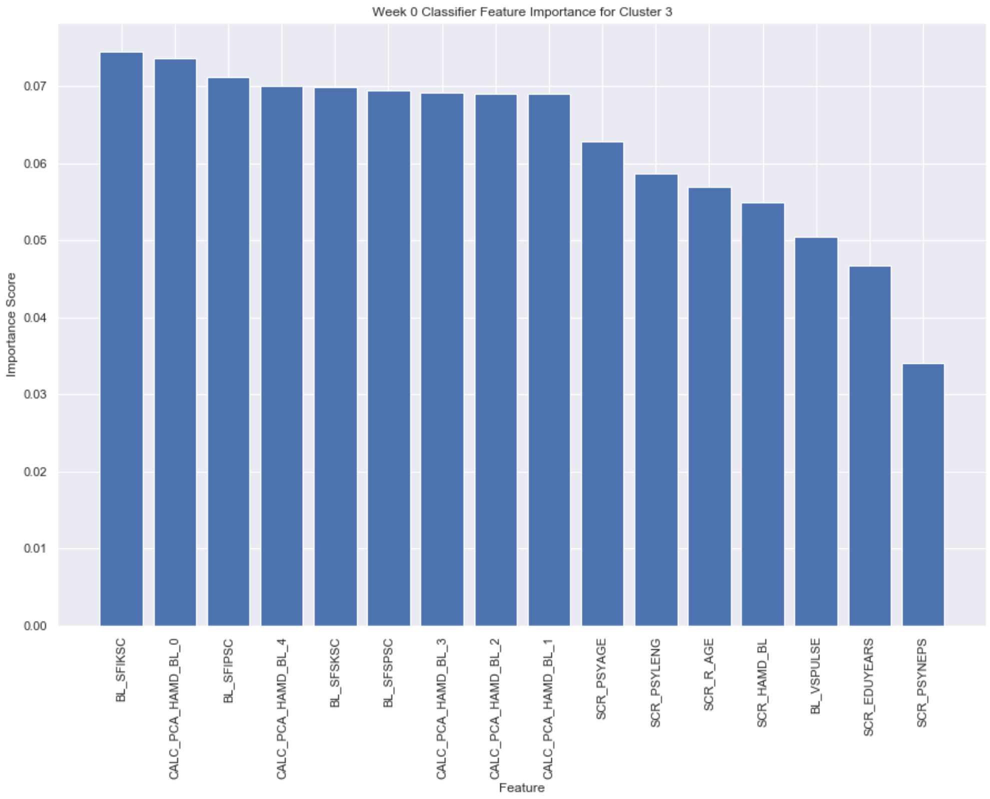

IDS variables - Week 3 Classifier Feature Importance for Response after 4 weeks
Importance Score
0.00
0.01
0.02
0.03
0.04
0.05
0.06
Feature
SCR_IDSCV03_rel
SCR_IDSFV03_rel
SCR_IDSCV03
SCR_IDSFV03
CALC_DELTA_IDSC_V03_BL
SCR_IDSCV02_rel
CALC_DELTA_IDSF_V03_BL
SCR_IDSCV02
CALC_PCA_IDSF_V02_0
CALC_DELTA_IDSC_V02_BL
CALC_PCA_IDSC_V02_1
CALC_PCA_IDSC_V02_0
SCR_IDSCV01_rel
SCR_IDSFV02_rel
SCR_IDSFV02
CALC_DELTA_IDSF_V03_V01
CALC_PCA_IDSF_V01_0
CALC_DELTA_IDSF_V03_V02
SCR_IDSFV01_rel
CALC_DELTA_IDSC_V03_V01
SCR_IDSCV01
CALC_PCA_IDSC_V02_5
SCR_IDSFV01
CALC_DELTA_IDSC_V01_BL
CALC_PCA_IDSC_BL_4
CALC_PCA_IDSC_V02_4
CALC_PCA_IDSC_V01_0
CALC_PCA_IDSF_BL_1
CALC_PCA_IDSC_V01_6
CALC_DELTA_IDSC_V03_V02

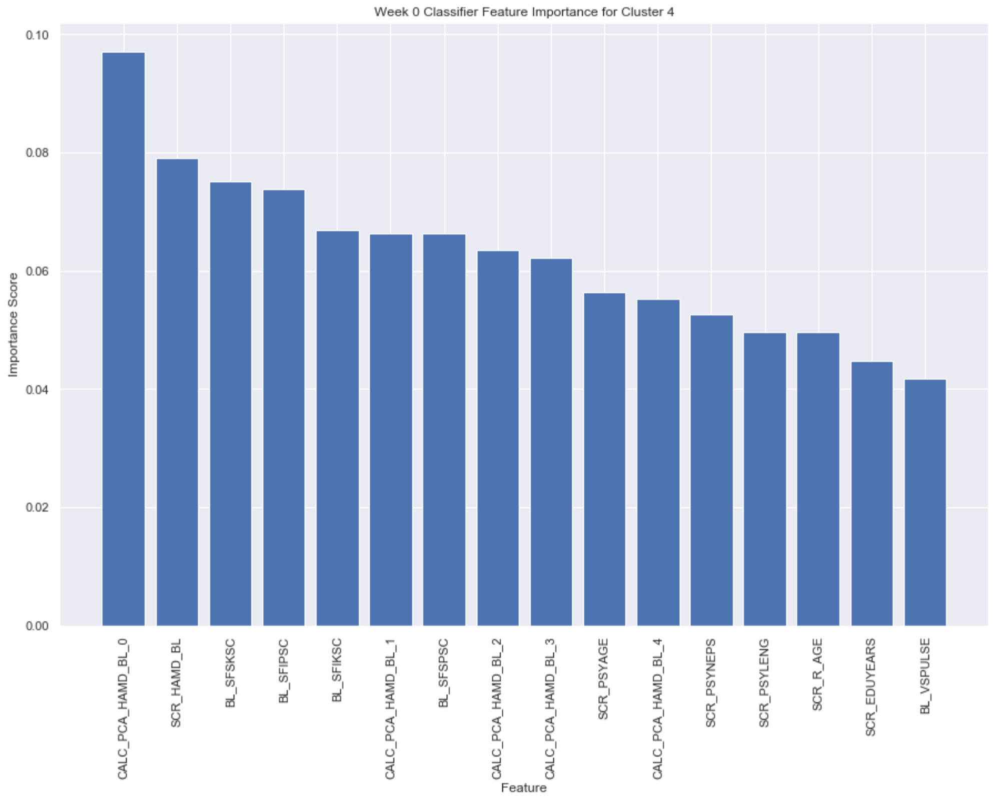
Experiment 3

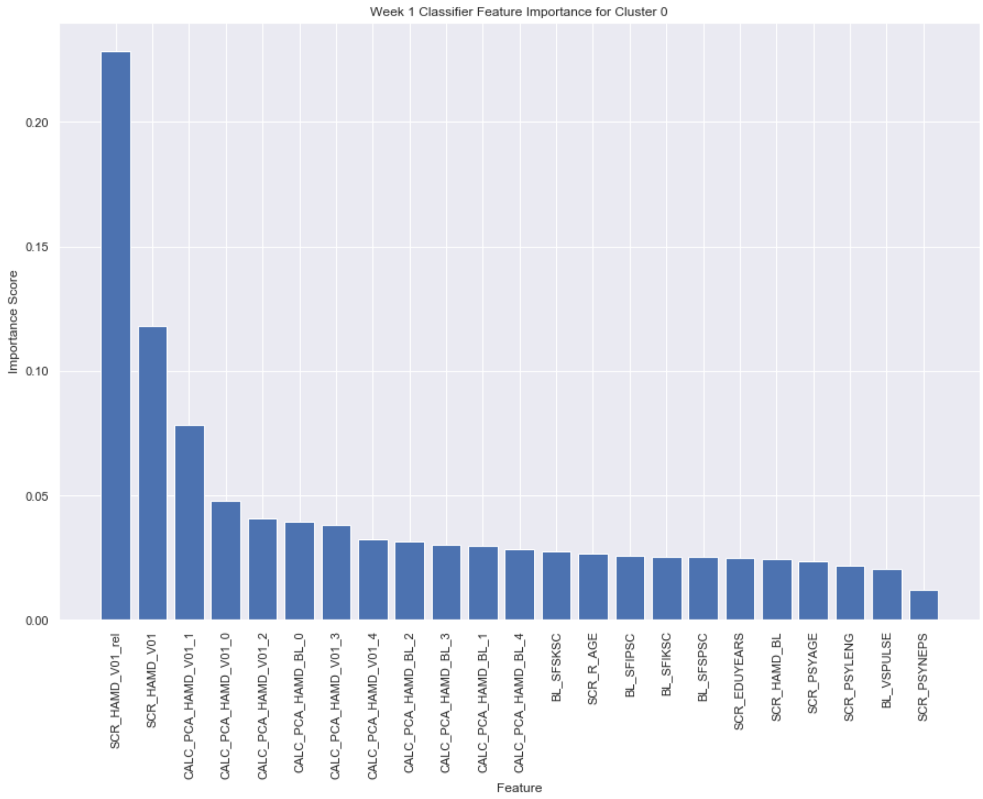


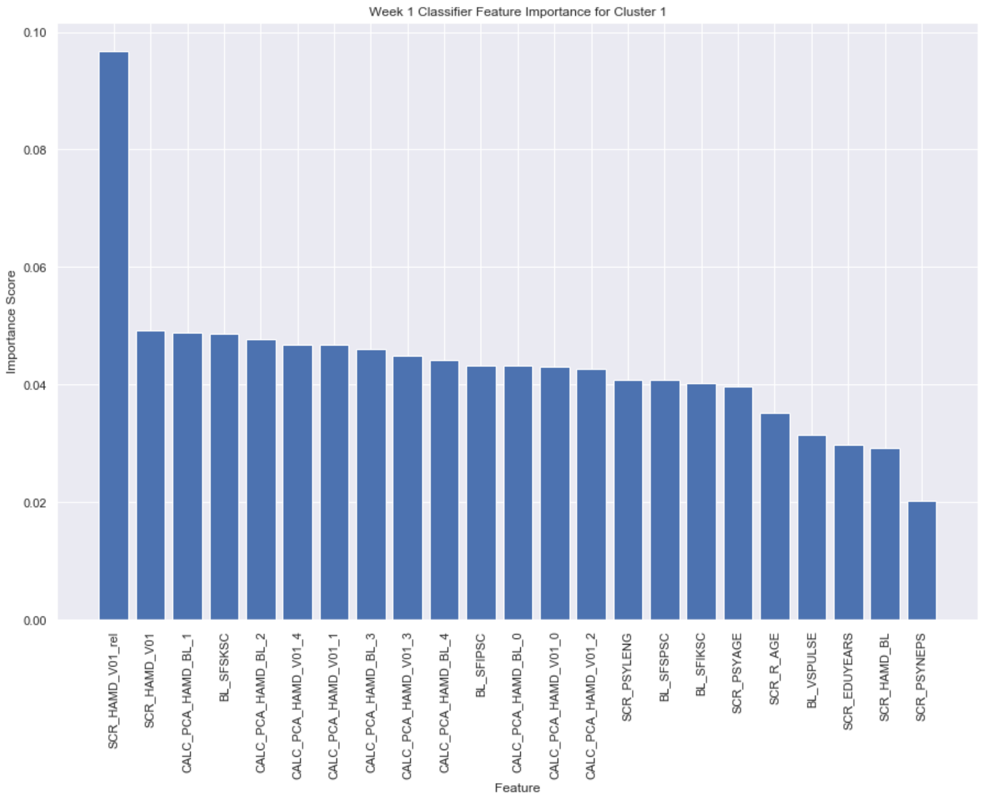


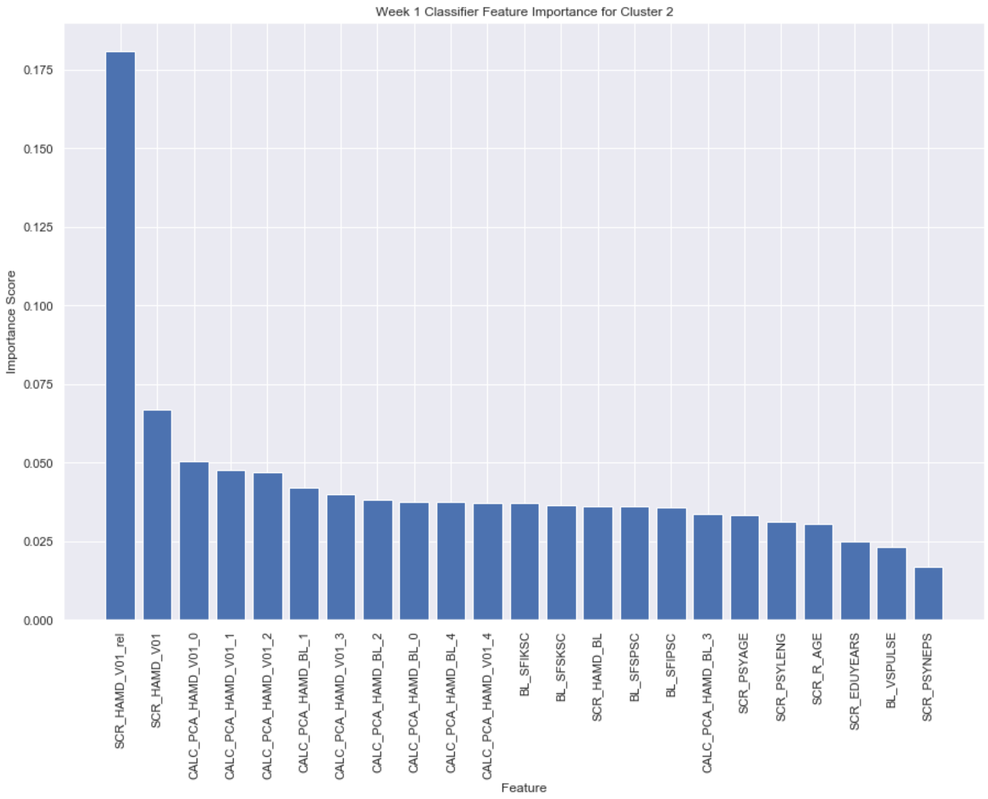

Week 0 Classifier Feature Importance for Cluster 2
Importance Score
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
CALC_PCA_HAMD_BL_1
BL_SFSKSC
CALC_PCA_HAMD_BL_2
CALC_PCA_HAMD_BL_0
BL_SFSPSC
CALC_PCA_HAMD_BL_3
CALC_PCA_HAMD_BL_4
SCR_PSYAGE
BL_SFIPSC
BL_SFIKSC
SCR_R_AGE
SCR_PSYLENG
SCR_HAMD_BL
SCR_EDUYEARS
BL_VSPULSE
SCR_PSYNEPS
Feature

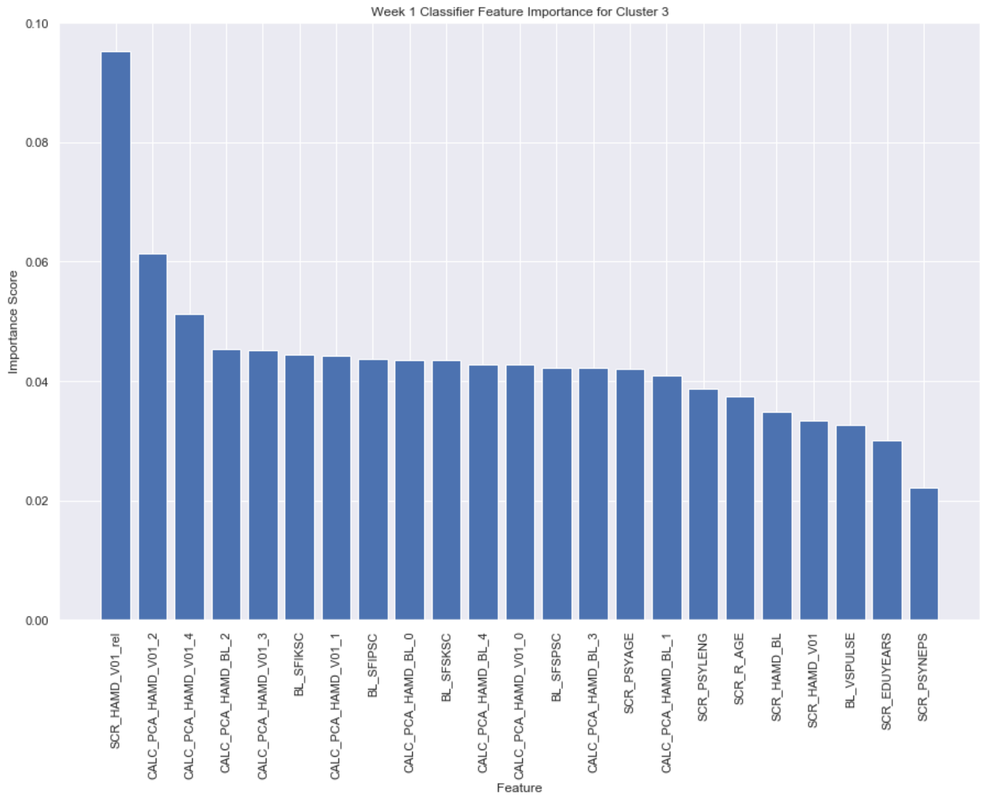

Week 0 Classifier Feature Importance for Cluster 3
Importance Score
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
BL_SFIKSC
CALC_PCA_HAMD_BL_0
BL_SFIPSC
CALC_PCA_HAMD_BL_4
BL_SFSKSC
BL_SFSPSC
CALC_PCA_HAMD_BL_3
CALC_PCA_HAMD_BL_2
CALC_PCA_HAMD_BL_1
SCR_PSYAGE
SCR_PSYLENG
SCR_R_AGE
SCR_HAMD_BL
BL_VSPULSE
SCR_EDUYEARS
SCR_PSYNEPS
Feature

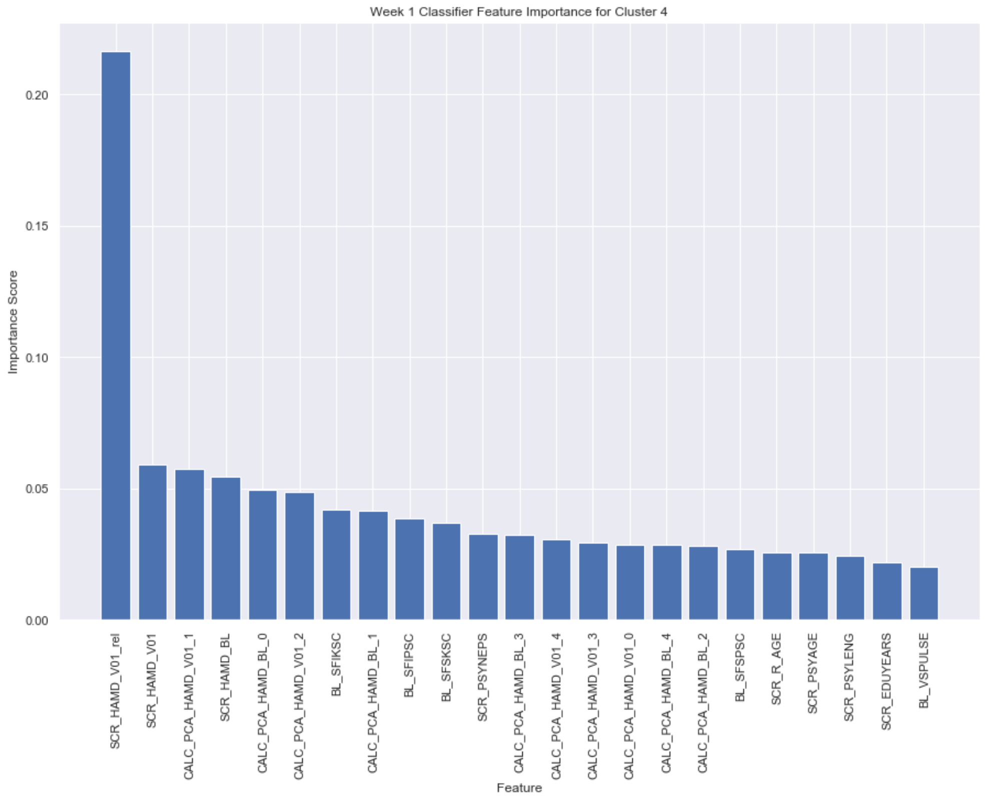

Week 0 Classifier Feature Importance for Cluster 4
Importance Score
0.10
0.08
0.06
0.04
0.02
0.00
CALC_PCA_HAMD_BL_0
SCR_HAMD_BL
BL_SFSKSC
BL_SFIPSC
BL_SFIKSC
CALC_PCA_HAMD_BL_1
BL_SFSPSC
CALC_PCA_HAMD_BL_2
CALC_PCA_HAMD_BL_3
SCR_PSYAGE
CALC_PCA_HAMD_BL_4
SCR_PSYNEPS
SCR_PSYLENG
SCR_R_AGE
SCR_EDUYEARS
BL_VSPULSE
Feature

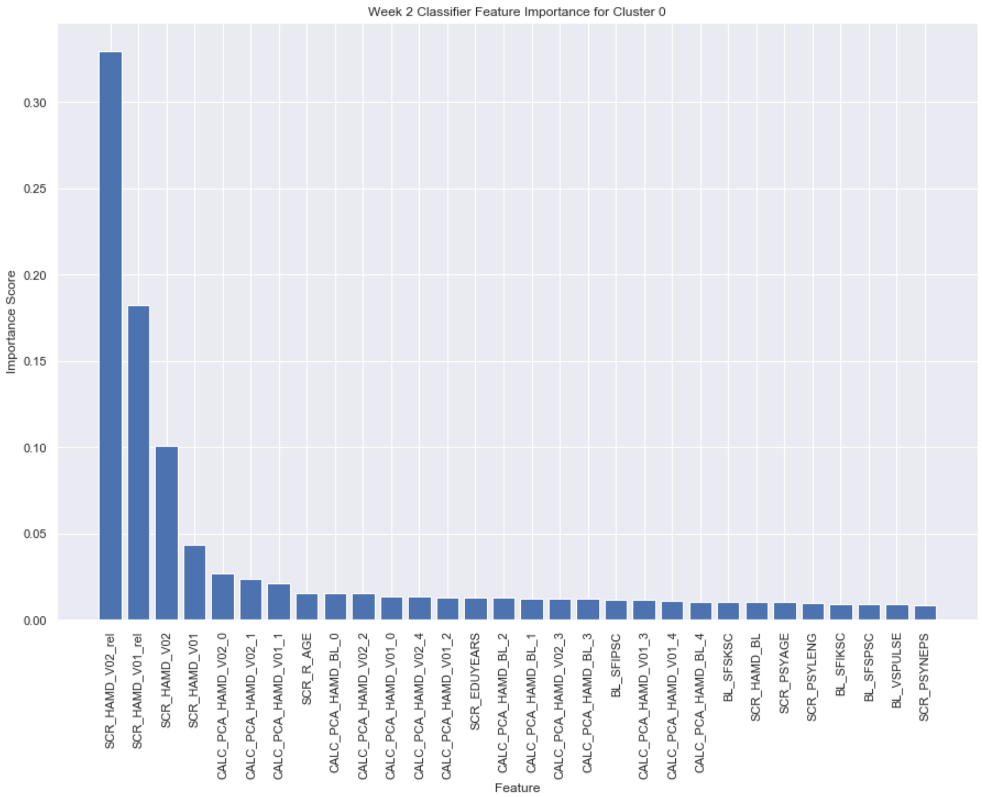

Week 1 Classifier Feature Importance for Cluster 0
Importance Score
0.20
0.15
0.10
0.05
0.00
SCR_HAMD_V01_rel
SCR_HAMD_V01
CALC_PCA_HAMD_V01_1
CALC_PCA_HAMD_V01_0
CALC_PCA_HAMD_V01_2
CALC_PCA_HAMD_BL_0
CALC_PCA_HAMD_V01_3
CALC_PCA_HAMD_V01_4
CALC_PCA_HAMD_BL_2
CALC_PCA_HAMD_BL_3
CALC_PCA_HAMD_BL_1
CALC_PCA_HAMD_BL_4
BL_SFSKSC
SCR_R_AGE
BL_SFIPSC
BL_SFIKSC
BL_SFSPSC
SCR_EDUYEARS
SCR_HAMD_BL
SCR_PSYAGE
SCR_PSYLENG
BL_VSPULSE
SCR_PSYNEPS
Feature

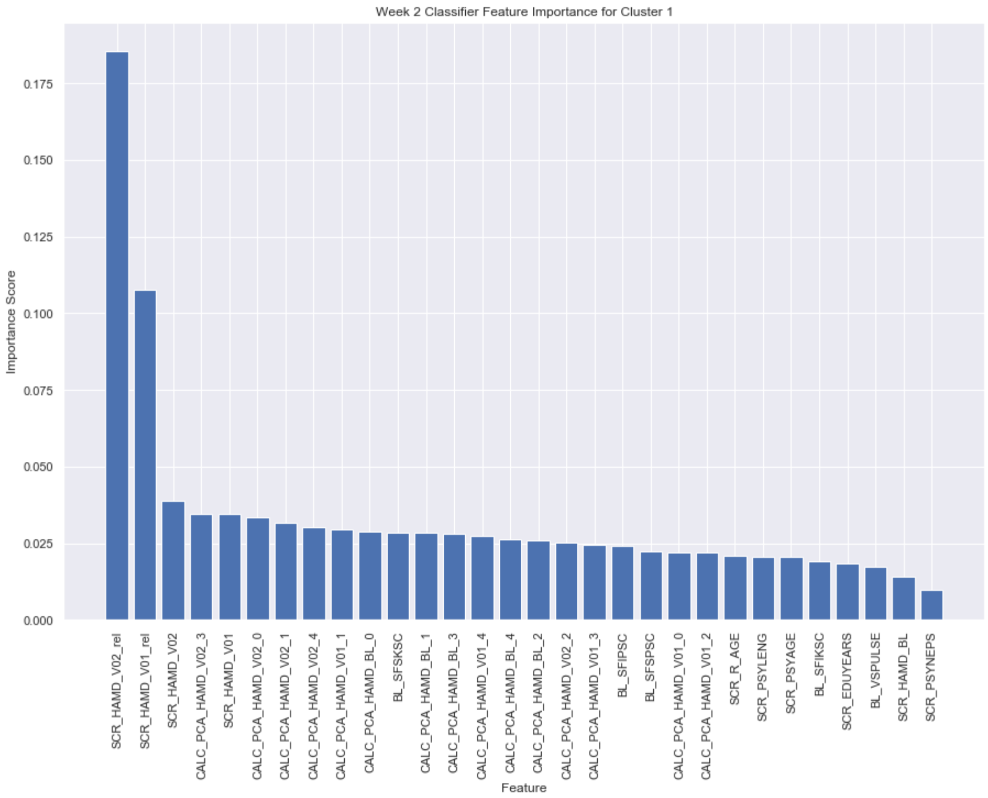

Week 1 Classifier Feature Importance for Cluster 1
Importance Score
0.10
0.08
0.06
0.04
0.02
0.00
SCR_HAMD_V01_rel
SCR_HAMD_V01
CALC_PCA_HAMD_BL_1
BL_SFSKSC
CALC_PCA_HAMD_BL_2
CALC_PCA_HAMD_V01_4
CALC_PCA_HAMD_V01_1
CALC_PCA_HAMD_BL_3
CALC_PCA_HAMD_V01_3
CALC_PCA_HAMD_BL_4
BL_SFIPSC
CALC_PCA_HAMD_BL_0
CALC_PCA_HAMD_V01_0
CALC_PCA_HAMD_V01_2
SCR_PSYLENG
BL_SFSPSC
BL_SFIKSC
SCR_PSYAGE
SCR_R_AGE
BL_VSPULSE
SCR_EDUYEARS
SCR_HAMD_BL
SCR_PSYNEPS
Feature

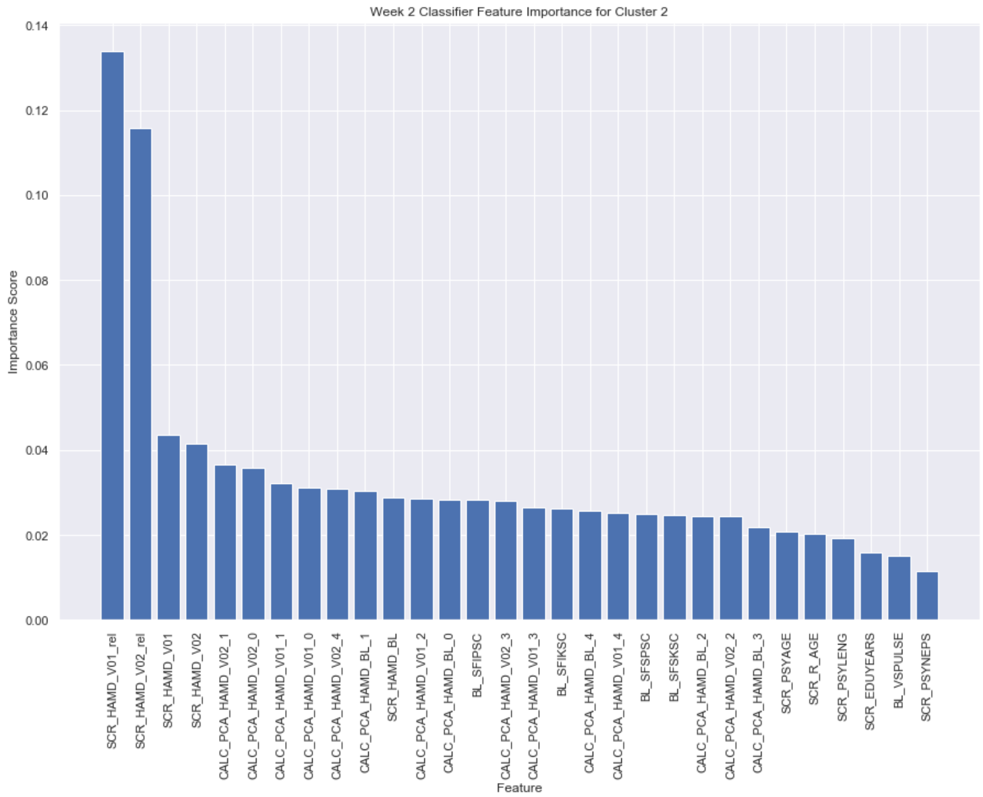

Week 1 Classifier Feature Importance for Cluster 2
Importance Score
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
SCR_HAMD_V01_rel
SCR_HAMD_V01
CALC_PCA_HAMD_V01_0
CALC_PCA_HAMD_V01_1
CALC_PCA_HAMD_V01_2
CALC_PCA_HAMD_BL_1
CALC_PCA_HAMD_V01_3
CALC_PCA_HAMD_BL_2
CALC_PCA_HAMD_BL_0
CALC_PCA_HAMD_BL_4
CALC_PCA_HAMD_V01_4
BL_SFIKSC
BL_SFSKSC
SCR_HAMD_BL
BL_SFSPSC
BL_SFIPSC
CALC_PCA_HAMD_BL_3
SCR_PSYAGE
SCR_PSYLENG
SCR_R_AGE
SCR_EDUYEARS
BL_VSPULSE
SCR_PSYNEPS
Feature


Week 1 Classifier Feature Importance for Cluster 3
Importance Score
0.10
0.08
0.06
0.04
0.02
0.00
Feature
SCR_HAMD_V01_rel
CALC_PCA_HAMD_V01_2
CALC_PCA_HAMD_V01_4
CALC_PCA_HAMD_BL_2
CALC_PCA_HAMD_V01_3
BL_SFIKSC
CALC_PCA_HAMD_V01_1
BL_SFIPSC
CALC_PCA_HAMD_BL_0
BL_SFSKSC
CALC_PCA_HAMD_BL_4
CALC_PCA_HAMD_V01_0
BL_SFSPSC
CALC_PCA_HAMD_BL_3
SCR_PSYAGE
CALC_PCA_HAMD_BL_1
SCR_PSYLENG
SCR_R_AGE
SCR_HAMD_BL
SCR_HAMD_V01
BL_VSPULSE
SCR_EDUYEARS
SCR_PSYNEPS

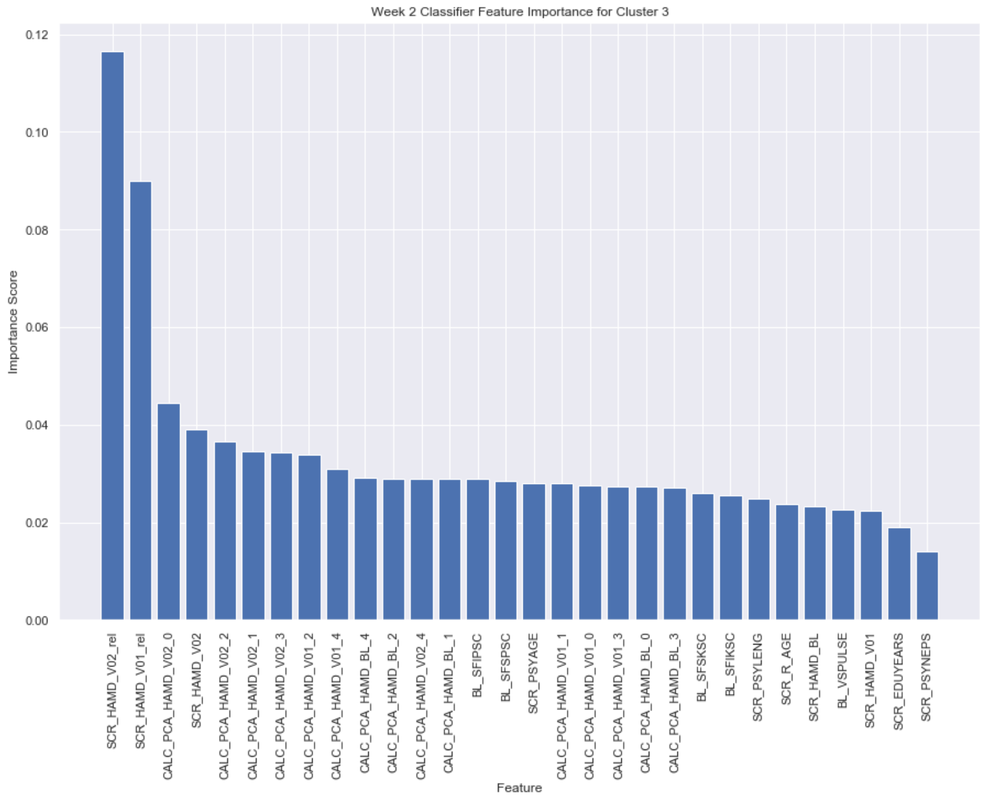

Week 1 Classifier Feature Importance for Cluster 4
Importance Score
0.20
0.15
0.10
0.05
0.00
Feature
SCR_HAMD_V01_rel
SCR_HAMD_V01
CALC_PCA_HAMD_V01_1
SCR_HAMD_BL
CALC_PCA_HAMD_BL_0
CALC_PCA_HAMD_V01_2
BL_SFIKSC
CALC_PCA_HAMD_BL_1
BL_SFIPSC
BL_SFSKSC
SCR_PSYNEPS
CALC_PCA_HAMD_BL_3
CALC_PCA_HAMD_V01_4
CALC_PCA_HAMD_V01_3
CALC_PCA_HAMD_V01_0
CALC_PCA_HAMD_BL_4
CALC_PCA_HAMD_BL_2
BL_SFSPSC
SCR_R_AGE
SCR_PSYAGE
SCR_PSYLENG
SCR_EDUYEARS
BL_VSPULSE


Week 2 Classifier Feature Importance for Cluster 0
Importance Score
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Feature
SCR_HAMD_V02_rel
SCR_HAMD_V01_rel
SCR_HAMD_V02
SCR_HAMD_V01
CALC_PCA_HAMD_V02_0
CALC_PCA_HAMD_V02_1
CALC_PCA_HAMD_V01_1
SCR_R_AGE
CALC_PCA_HAMD_BL_0
CALC_PCA_HAMD_V02_2
CALC_PCA_HAMD_V01_0
CALC_PCA_HAMD_V02_4
CALC_PCA_HAMD_V01_2
SCR_EDUYEARS
CALC_PCA_HAMD_BL_2
CALC_PCA_HAMD_BL_1
CALC_PCA_HAMD_V02_3
CALC_PCA_HAMD_BL_3
BL_SFIPSC
CALC_PCA_HAMD_V01_3
CALC_PCA_HAMD_V01_4
CALC_PCA_HAMD_BL_4
BL_SFSKSC
SCR_HAMD_BL
SCR_PSYAGE
SCR_PSYLENG
BL_SFIKSC
BL_SFSPSC
BL_VSPULSE
SCR_PSYNEPS

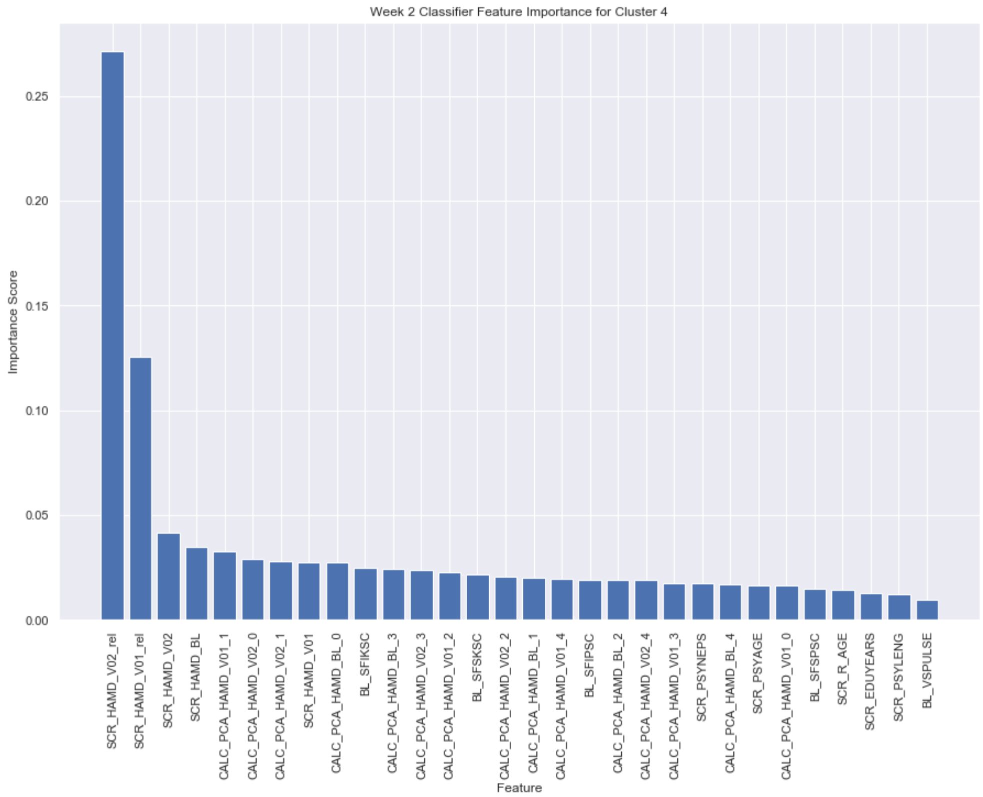

Week 2 Classifier Feature Importance for Cluster 1
Importance Score
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
Feature
SCR_HAMD_V02_rel
SCR_HAMD_V01_rel
SCR_HAMD_V02
CALC_PCA_HAMD_V02_3
SCR_HAMD_V01
CALC_PCA_HAMD_V02_0
CALC_PCA_HAMD_V02_1
CALC_PCA_HAMD_V02_4
CALC_PCA_HAMD_V01_1
CALC_PCA_HAMD_BL_0
BL_SFSKSC
CALC_PCA_HAMD_BL_1
CALC_PCA_HAMD_BL_3
CALC_PCA_HAMD_V01_4
CALC_PCA_HAMD_BL_4
CALC_PCA_HAMD_BL_2
CALC_PCA_HAMD_V02_2
CALC_PCA_HAMD_V01_3
BL_SFIPSC
BL_SFSPSC
CALC_PCA_HAMD_V01_0
CALC_PCA_HAMD_V01_2
SCR_R_AGE
SCR_PSYLENG
SCR_PSYAGE
BL_SFIKSC
SCR_EDUYEARS
BL_VSPULSE
SCR_HAMD_BL
SCR_PSYNEPS

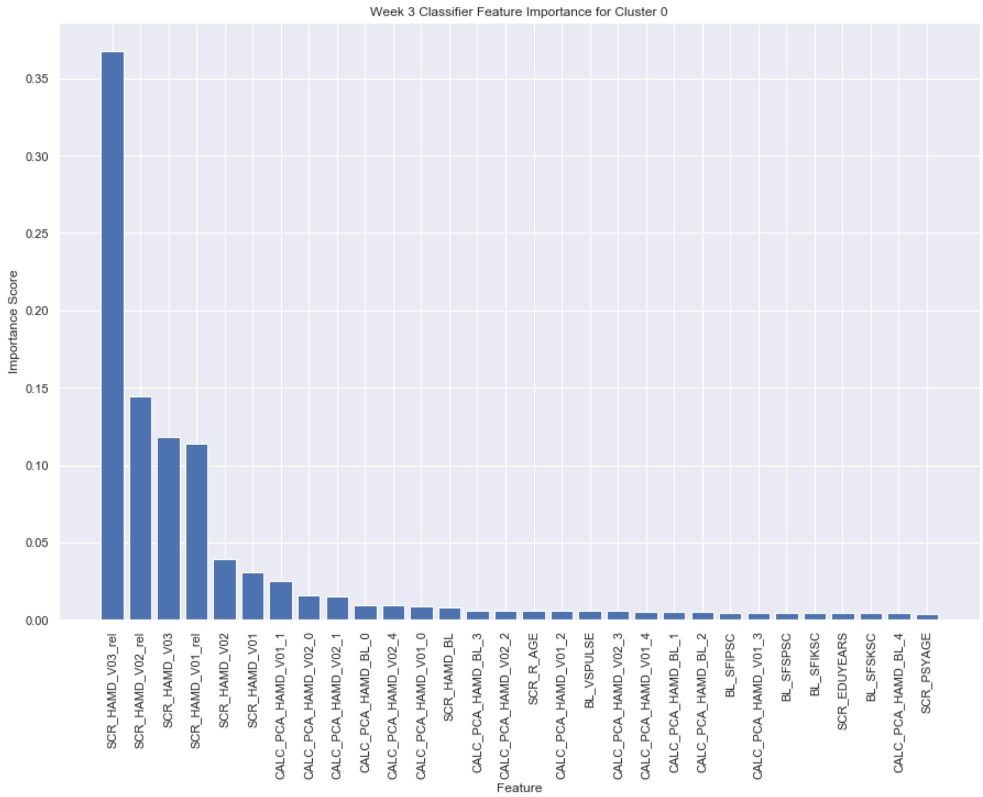


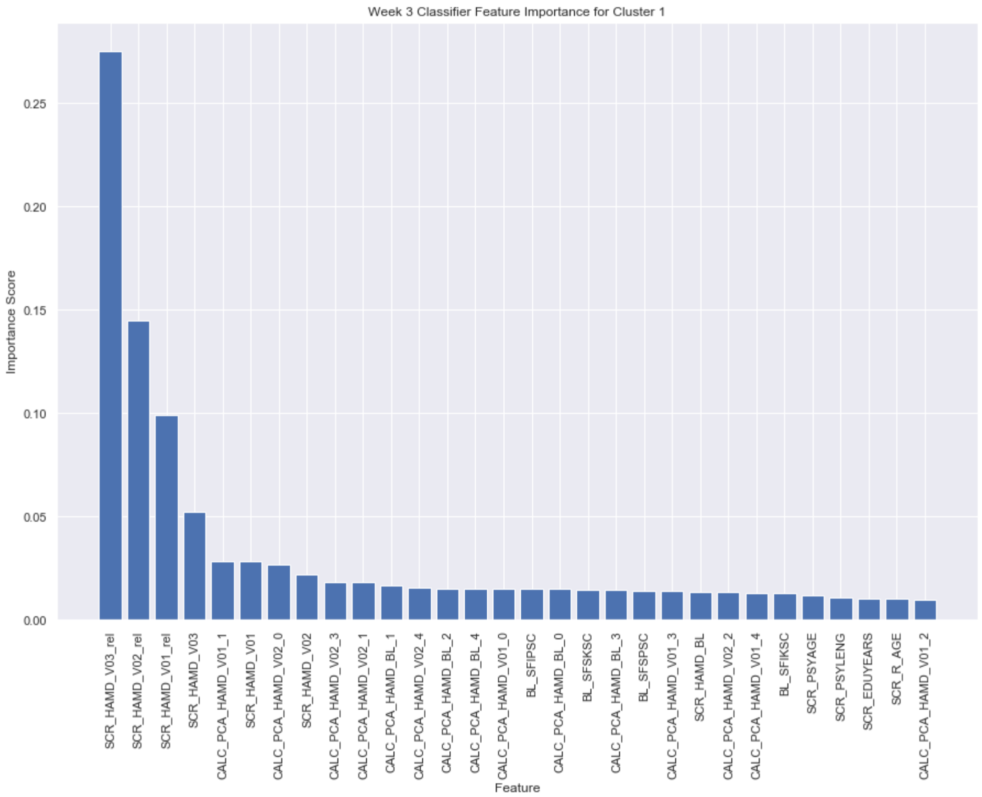
Week 2 Classifier Feature Importance for Cluster 2

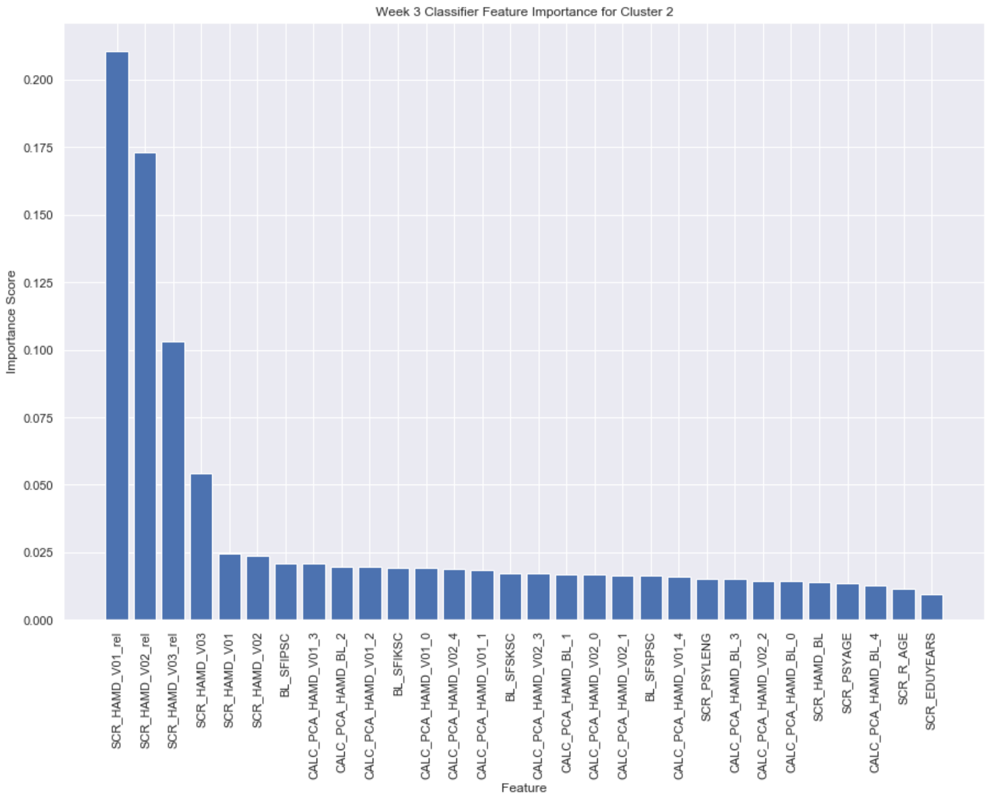


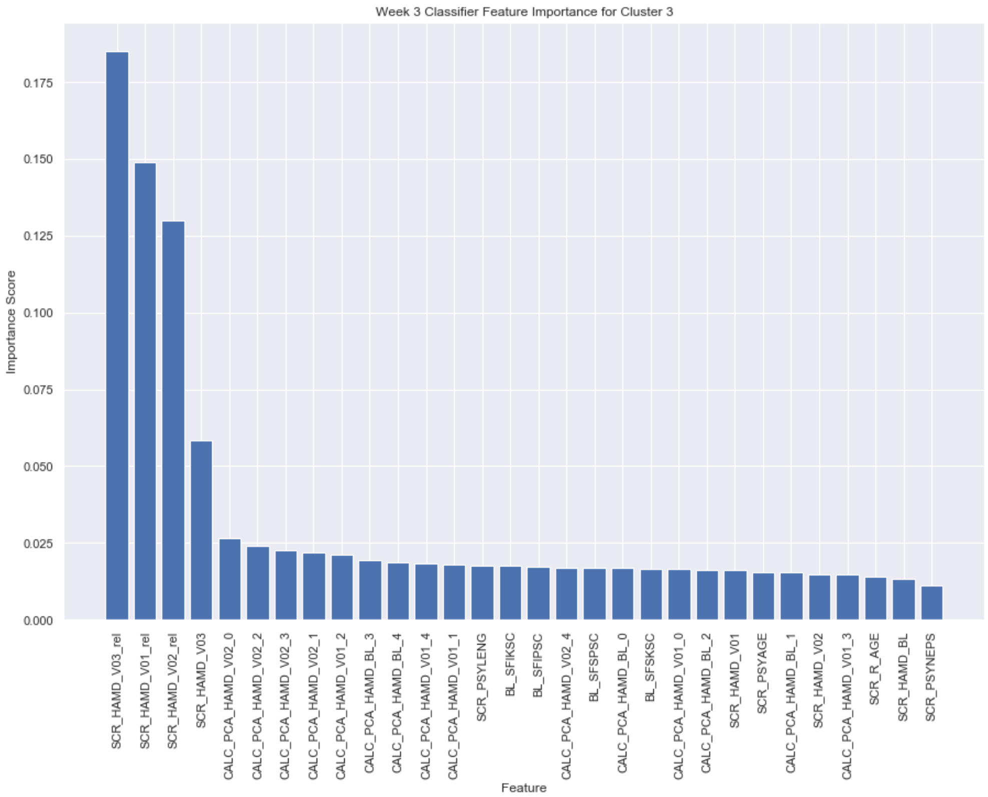
Week 2 Classifier Feature Importance for Cluster 3

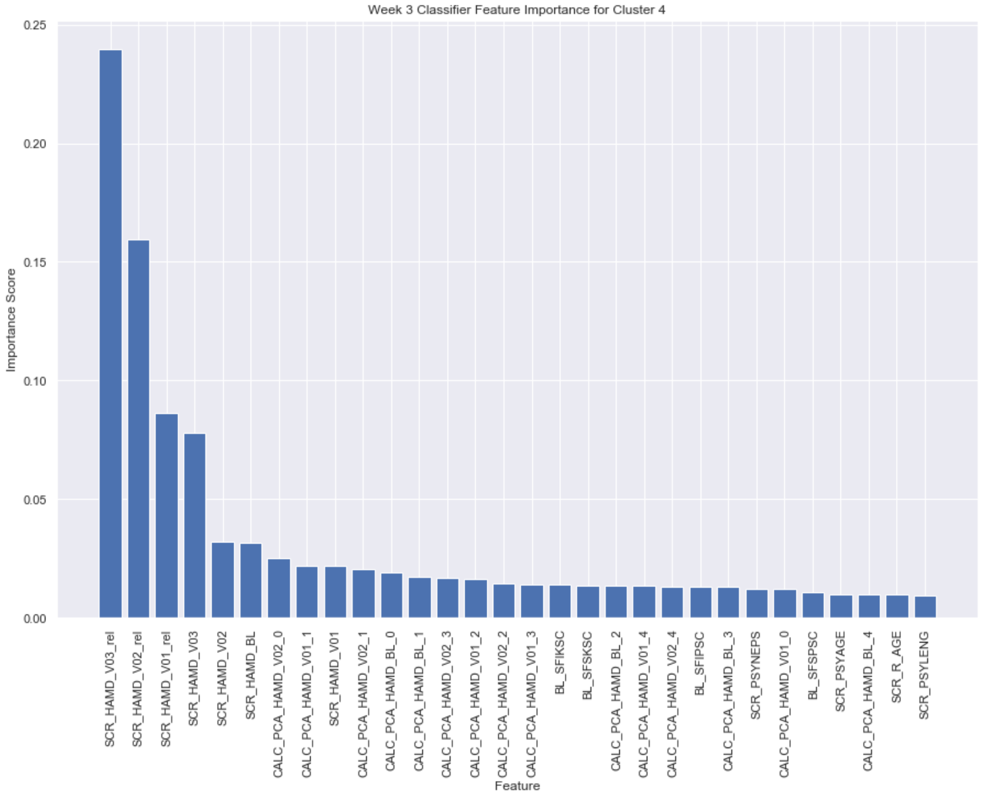

Week 2 Classifier Feature Importance for Cluster 4
Importance Score
0.25
0.20
0.15
0.10
0.05
0.00
SCR_HAMD_V02_rel
SCR_HAMD_V01_rel
SCR_HAMD_V02
SCR_HAMD_BL
CALC_PCA_HAMD_V01_1
CALC_PCA_HAMD_V02_0
CALC_PCA_HAMD_V02_1
SCR_HAMD_V01
CALC_PCA_HAMD_BL_0
BL_SFIKSC
CALC_PCA_HAMD_BL_3
CALC_PCA_HAMD_V02_3
CALC_PCA_HAMD_V01_2
BL_SFSKSC
CALC_PCA_HAMD_V02_2
CALC_PCA_HAMD_BL_1
CALC_PCA_HAMD_V01_4
BL_SFIPSC
CALC_PCA_HAMD_BL_2
CALC_PCA_HAMD_V02_4
CALC_PCA_HAMD_V01_3
SCR_PSYNEPS
CALC_PCA_HAMD_BL_4
SCR_PSYAGE
CALC_PCA_HAMD_V01_0
BL_SFSPSC
SCR_R_AGE
SCR_EDUYEARS
SCR_PSYLENG
BL_VSPULSE
Feature


Week 3 Classifier Feature Importance for Cluster 0
Importance Score
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
SCR_HAMD_V03_rel
SCR_HAMD_V02_rel
SCR_HAMD_V03
SCR_HAMD_V01_rel
SCR_HAMD_V02
SCR_HAMD_V01
CALC_PCA_HAMD_V01_1
CALC_PCA_HAMD_V02_0
CALC_PCA_HAMD_V02_1
CALC_PCA_HAMD_BL_0
CALC_PCA_HAMD_V02_4
CALC_PCA_HAMD_V01_0
SCR_HAMD_BL
CALC_PCA_HAMD_BL_3
CALC_PCA_HAMD_V02_2
SCR_R_AGE
CALC_PCA_HAMD_V01_2
BL_VSPULSE
CALC_PCA_HAMD_V02_3
CALC_PCA_HAMD_V01_4
CALC_PCA_HAMD_BL_1
CALC_PCA_HAMD_BL_2
BL_SFIPSC
CALC_PCA_HAMD_V01_3
BL_SFSPSC
BL_SFIKSC
SCR_EDUYEARS
BL_SFSKSC
CALC_PCA_HAMD_BL_4
SCR_PSYAGE
Feature


Week 3 Classifier Feature Importance for Cluster 1
Importance Score
0.25
0.20
0.15
0.10
0.05
0.00
SCR_HAMD_V03_rel
SCR_HAMD_V02_rel
SCR_HAMD_V01_rel
SCR_HAMD_V03
CALC_PCA_HAMD_V01_1
SCR_HAMD_V01
CALC_PCA_HAMD_V02_0
SCR_HAMD_V02
CALC_PCA_HAMD_V02_3
CALC_PCA_HAMD_V02_1
CALC_PCA_HAMD_BL_1
CALC_PCA_HAMD_V02_4
CALC_PCA_HAMD_BL_2
CALC_PCA_HAMD_BL_4
CALC_PCA_HAMD_V01_0
BL_SFIPSC
CALC_PCA_HAMD_BL_0
BL_SFSKSC
CALC_PCA_HAMD_BL_3
BL_SFSPSC
CALC_PCA_HAMD_V01_3
SCR_HAMD_BL
CALC_PCA_HAMD_V02_2
CALC_PCA_HAMD_V01_4
BL_SFIKSC
SCR_PSYAGE
SCR_PSYLENG
SCR_EDUYEARS
SCR_R_AGE
CALC_PCA_HAMD_V01_2
Feature


Week 3 Classifier Feature Importance for Cluster 2
Importance Score
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
SCR_HAMD_V01_rel
SCR_HAMD_V02_rel
SCR_HAMD_V03_rel
SCR_HAMD_V03
SCR_HAMD_V01
SCR_HAMD_V02
BL_SFIPSC
CALC_PCA_HAMD_V01_3
CALC_PCA_HAMD_BL_2
CALC_PCA_HAMD_V01_2
BL_SFIKSC
CALC_PCA_HAMD_V01_0
CALC_PCA_HAMD_V02_4
CALC_PCA_HAMD_V01_1
BL_SFSKSC
CALC_PCA_HAMD_V02_3
CALC_PCA_HAMD_BL_1
CALC_PCA_HAMD_V02_0
CALC_PCA_HAMD_V02_1
BL_SFSPSC
CALC_PCA_HAMD_V01_4
SCR_PSYLENG
CALC_PCA_HAMD_BL_3
CALC_PCA_HAMD_V02_2
CALC_PCA_HAMD_BL_0
SCR_HAMD_BL
SCR_PSYAGE
CALC_PCA_HAMD_BL_4
SCR_R_AGE
SCR_EDUYEARS
Feature


Week 3 Classifier Feature Importance for Cluster 3
Importance Score
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
Feature
SCR_HAMD_V03_rel
SCR_HAMD_V01_rel
SCR_HAMD_V02_rel
SCR_HAMD_V03
CALC_PCA_HAMD_V02_0
CALC_PCA_HAMD_V02_2
CALC_PCA_HAMD_V02_3
CALC_PCA_HAMD_V02_1
CALC_PCA_HAMD_V01_2
CALC_PCA_HAMD_BL_3
CALC_PCA_HAMD_BL_4
CALC_PCA_HAMD_V01_4
CALC_PCA_HAMD_V01_1
SCR_PSYLENG
BL_SFIKSC
BL_SFIPSC
CALC_PCA_HAMD_V02_4
BL_SFSPSC
CALC_PCA_HAMD_BL_0
BL_SFSKSC
CALC_PCA_HAMD_V01_0
CALC_PCA_HAMD_BL_2
SCR_HAMD_V01
SCR_PSYAGE
CALC_PCA_HAMD_BL_1
SCR_HAMD_V02
CALC_PCA_HAMD_V01_3
SCR_R_AGE
SCR_HAMD_BL
SCR_PSYNEPS


Week 3 Classifier Feature Importance for Cluster 4
Importance Score
0.00
0.05
0.10
0.15
0.20
0.25
Feature
SCR_HAMD_V03_rel
SCR_HAMD_V02_rel
SCR_HAMD_V01_rel
SCR_HAMD_V03
SCR_HAMD_V02
SCR_HAMD_BL
CALC_PCA_HAMD_V02_0
CALC_PCA_HAMD_V01_1
SCR_HAMD_V01
CALC_PCA_HAMD_V02_1
CALC_PCA_HAMD_BL_0
CALC_PCA_HAMD_BL_1
CALC_PCA_HAMD_V02_3
CALC_PCA_HAMD_V01_2
CALC_PCA_HAMD_V02_2
CALC_PCA_HAMD_V01_3
BL_SFIKSC
BL_SFSKSC
CALC_PCA_HAMD_BL_2
CALC_PCA_HAMD_V01_4
CALC_PCA_HAMD_V02_4
BL_SFIPSC
CALC_PCA_HAMD_BL_3
SCR_PSYNEPS
CALC_PCA_HAMD_V01_0
BL_SFSPSC
SCR_PSYAGE
CALC_PCA_HAMD_BL_4
SCR_R_AGE
SCR_PSYLENG

Acknowledgement

Not available in the pdf version

Curriculum vitae

Not available in the pdf version