Deconstructing depression: unified syndrome or groups of symptoms?

Eeden, W.A. van

Chapter 7

Predicting the 9-year course of mood and anxiety disorders with automated machine learning: A comparison between auto-sklearn, naïve Bayes classifier, and traditional logistic regression

van Eeden, W.A.,, Chuan, L., van Hemert, A.M., Carlier, I.V.E. , Penninx, B.W.c, Wardenaar, K.J., Hoos, H., Giltay, E.J.

(2021). Psychiatry Research, 299, 113823.

Abstract

Background: Predicting the onset and course of mood and anxiety disorders is of clinical importance but remains difficult. We compared the predictive performances of traditional logistic regression, basic probabilistic machine learning (ML) methods, and automated ML (Auto-sklearn).

Methods: Data were derived from the Netherlands Study of Depression and Anxiety. We compared how well multinomial logistic regression, a naïve Bayes classifier, and Auto-sklearn predicted depression and anxiety diagnoses at a 2-, 4-, 6-, and 9-year follow up, operationalized as binary or categorical variables. Predictor sets included demographic and self-report data, which can be easily collected in clinical practice at two initial time points (baseline and 1-year follow up).

Results: At baseline, participants were 42.2 years old, 66.5% were women, and 53.6% had a current mood or anxiety disorder. The three methods were similarly successful in predicting (mental) health status, with correct predictions for up to 79% (95% CI 75–81%). However, Auto-sklearn was superior when assessing a more complex dataset with individual item scores.

Conclusions: Automated ML methods added only limited value, compared to traditional data modelling when predicting the onset and course of depression and anxiety. However, they hold potential for automatization and may be better suited for complex datasets.

Keywords: Psychiatry, Depression, Anxiety disorder, Machine Learning, Logistic Models, Epidemiologic Methods, Regression Analysis

Highlights

- The predictive performances were compared between a automated machine learning algorithm, a basic probabilistic ML algorithm and more traditional multinominal logistic regression when predicting depression and anxiety at 2-, 4-, 6-, 9-year follow-up.
- In 96 models, we used multiple sets of demographic and self-report questionnaire data as predictor variables, which can be easily collected in clinical practice at two initial time points (baseline and 1-year follow up).
- Depression and anxiety could be predicted with correct predictions of up to 79%.
- None of the methods seemed to consistently outperform one another. Although, Auto-sklearn was superior when using a more complex data-set with individual item- scores.
- Clinical practice as may in time benefit from integrating next generation ML methods into clinical discussion making due to its potential for automatization and its adaptability for more complex datasets, rather than its increased predictive accuracy compared to more traditional data modelling methods.

7.1 Introduction

Despite a large body of epidemiological research, the course and onset of mood and anxiety disorders remain difficult to predict. Improving the ability to predict the onset and course of mood and anxiety disorders can be clinically relevant for prevention, early detection, staging, and personalized treatments [1]. In clinical settings, most decision making is based on clinical-care guidelines and experience [2]. However, even experienced clinicians may ignore relevant information or may put too much emphasis on clinically salient cues [3]. Information on demographic characteristics and clinician-rated and self-reported measures are increasingly collected as part of routine outcome monitoring [ROM; 4], but this information is underused in clinical decision making. Literature suggests that automated statistical prediction of current diagnoses and course may improve clinical decision making [2, 5], particularly through modern machine learning (ML) approaches [6].

ML may be more time efficient, better suited for large and complex datasets, and better able to detect complex patterns in the data than current data-modelling approaches that rely heavily on human decision making [7, 8]. Most clinical data thus far have been analyzed by selecting only specific putative predictors. It is possible that more complex (including nonlinear and higher dimensional) patterns exist in the data, which can efficiently be detected when analyzing all available data simultaneously using ML [9, 10]. These approaches are able to examine huge numbers of potential predictors in an unbiased manner while preventing overfitting [11].

Thus far, ML studies in the field of psychiatry have been promising. A recent meta-analysis, which included 20 studies that predicted the therapeutic outcome of depression using ML algorithms, found an overall accuracy of .82 [95% CI 0.77–0.87; 12]. Another ML study used an extensive set of baseline variables in a subset of 805 depressed patients from the Netherlands Study of Depression and Anxiety (NESDA) cohort, including biological and psychological variables [e.g., personality traits; 13]. The study achieved an accuracy significantly greater than chance of 66% for predicting persistent depression over the course of 2 years. A similar study, performed in a subset of the NESDA cohort of 887 anxiety patients, found an accuracy of predicting anxiety recovery of 62% ($p < .05$) and an accuracy of predicting recovery of all common mental disorders of 63% [$p < 0.05$; 14]. Clinical severity measures were the most important predictor variables, which is in line with previous reports [12-14]. Although these studies seem promising, recently published papers have demonstrated only limited added value of ML over traditional regression analyzes [15, 16].

Additionally, other studies found that when predicting suicide, ML did not outperform regression analysis and resulted in positive predictive values below 0.01, thus limiting the practical utility of these predictions [17, 18]. Despite the increasing number of publications in this field, ML has yet to move towards clinical application [19].

Although ML incorporates less human decision making than traditional methods, most ML methods are still not fully automated. Feature selection has been standardized as much as possible, but cut-off values that determine which features to include or exclude are somewhat arbitrarily selected. One solution would be to fully automate the selection of features, as is done in the Auto-sklearn system [20]. Auto-sklearn is a next generation ML system that automatically selects the learning algorithm that best suits the data and automatically optimizes the hyperparameter settings of this algorithm. It has proved effective when analyzing a diverse range of datasets and is considered to be an efficient and robust system for use by both ML novices and experts [21, 22].

We aimed to study and to compare the performance of traditional multinominal logistic regression, a basic probabilistic ML algorithm [naïve Bayesian classifier; 23] and a more advanced automated ML method (Auto-sklearn) to predict DSM-IV-TR psychiatric diagnoses at a 2-, 4-, 6-, and 9-year follow up with different sets of predictors. We incorporated predictor variables that can be easily and inexpensively collected in clinical practice, such as demographic variables, clinician-rated psychiatric diagnoses, and self-reported depression and anxiety. Our hypothesis was that Auto-sklearn would be better at detecting complex patterns in the data and therefore would outdo a naïve Bayesian classifier, which in turn would outdo traditional regression analysis techniques in achieved level of accuracy. Moreover, we hypothesized that Auto-sklearn would be particularly efficient when single items and follow-up measures were included.

7.2 Methods

7.2.1 Study sample and procedures

For the current study, we included participants from the NESDA cohort, which investigated the course and consequences of depressive and anxiety disorders. A detailed description of the NESDA design and sampling procedures are published elsewhere [24]. The first wave (baseline) lasted from 2004 to September 2007, and the sixth wave of measurement at the 9-year follow up finished in October 2016. NESDA is a cohort study that recruited from the community (*n* = 564; 18.9%), general practice (*n* = 1,610; 54.0%), and secondary mental healthcare [n=807; 27.1%; 24] and included patients with a current or lifetime depressive or anxiety disorder as well as healthy controls (see supplementary Table 1). A limited number of exclusion criteria were applied, namely not being fluent in Dutch and the presence of other clinically overt psychiatric disorders (e.g., addiction, psychotic, bipolar). With this method, NESDA aimed for a cohort that is representative for diverse populations of healthy controls and patients with depression and anxiety [24]. Due to missing outcome data (mainly due to attrition), we included 2,596 (87.1%) participants to predict 2-year outcomes, 2,402 (80.6%) to predict 4-year outcomes, 2,256 (75.7%) to predict 6-year outcomes, and 2,068 (69.4%) to predict 9-year outcomes.

7.2.2 Measures

7.2.2.1 Independent variables

An overview of the independent variables within each predictor set can be found in Table 1 in the supplementary material. Independent variables comprised baseline demographics, lifetime and baseline DSM-IV-TR diagnoses, self-reported depression, and anxiety symptomatology. Demographic variables included gender, age, ethnicity (North European heritage: yes/no), level of education (1 = elementary or less; 2 = general intermediate/secondary education; 3 = college/university), partner status (no partner, with partner [not married], married, living apart/no partner, divorced/no partner, widowed/no partner), and working status (employed/unemployed). The Composite International Diagnostic Interview (CIDI WHO, version 2.1) was used to assess the presence of mood and anxiety disorders according to the DSM-IV-TR. This included current dysthymia, major depressive disorder (MDD), lifetime depressive disorder, social phobia, panic with agoraphobia, panic without agoraphobia, agoraphobia without panic, generalized anxiety disorder, and lifetime anxiety disorder. Future CIDI-based diagnoses were used as outcome variables at 2-, 4-, 6-, and 9-year follow up, and past and current CIDI-based diagnoses were used as independent

variables. Thus, diagnoses at baseline and at Years 2, 4, and 6 were used to predict the diagnosis at the 9-year follow up (see Section 2.2.2).

Anxiety and depressive severity as well as symptoms at baseline and 1-year follow up were assessed using the Fear Questionnaire [FQ; 25], the Beck's Anxiety Inventory [BAI; 26], and the Inventory of Depressive Symptomatology [IDS-SR; 27]. These measures were entered into the models as either sum scores only or as a combination of sum scores and individual items. Detailed (psychometric) information about the measures can be found in the supplementary material.

7.2.2.2 Outcome variable: Clinical diagnoses

The CIDI WHO, version 2.1 was used to assess clinical diagnoses according to the DSM-IV-TR. The CIDI is a fully standardized diagnostic interview with extensively validated psychometric characteristics [24, 28] and may be considered a gold standard for psychiatric diagnostic classification [29, 30].

At the 2-, 4-, 6-, and 9-year follow up, CIDI-based outcomes were coded both as a binary variable (psychiatric disorder absent vs. present) and as a categorical variable with four categories: healthy, mood disorder (i.e., major depression and/or dysthymia), anxiety disorder (i.e., general anxiety, social phobia, panic with agoraphobia, panic without agoraphobia, and/or agoraphobia without a panic disorder), and comorbid mood and anxiety disorders.

7.2.3 Statistical analysis

A total of 96 models were tested. We compared three methods, over four sets of predictor variables, over two outcome sets, and over four follow-up waves. The three methods were multinomial logistic regression [31], naïve Bayes classifier [23], and Auto-sklearn [21]. The four sets of predictor variables (all including sociodemographic variables and baseline diagnoses) were (a) baseline sum scores only; (b) baseline sum scores and 1-year follow up sum scores; (c) baseline sum scores, 1-year follow up sum scores, and individual items at baseline; and (d) sum scores and individual items at baseline and 1-year follow up. For an overview of the predictor Sets A–D, see Table 1 in the supplementary material. Missing item values (0.54% –13.1%) were replaced by the mean of the available cases. The two outcomes were binary (healthy/mood or anxiety disorder) and multinomial (healthy [A], mood disorder [B], anxiety [C], or comorbid mood- and anxiety disorder [D]). The follow-up waves occurred at 2, 4, 6, and 9 years.

Auto-sklearn is an automated ML system that addresses both the problem of choosing which ML algorithm is best suited to analyze a specific application scenario (i.e., the model/algorithm selection problem) and the problem of determining which parameter setting leads to high performance (i.e., the hyperparameter optimization problem). Auto-sklearn considers a wide range of feature selection methods including all classification approaches implemented within the Python `scikit-learn` package, spanning 15 classifiers (e.g., random forests, decision tree, gradient boosting, etc.), 14 feature preprocessing methods (e.g., feature agglomeration, polynomial, nystroem sampler, etc.), and four data preprocessing methods (i.e., one-hot encoding, imputation, balancing, and rescaling), giving rise to a structured hypothesis space with 110 hyperparameters. Auto-sklearn features preprocessing methods that can be mainly categorized into feature selection, kernel approximation, matrix decomposition, embeddings, feature clustering, polynomial feature expansion, and methods that use a classifier for feature selection [for more details see; 22]. Previous research shows that the classification performance is often much better than using standard selection/hyperparameter optimization methods [21], and researchers believe Auto-sklearn to be a promising system for use by both ML novices and experts [22]. Auto-sklearn won six out of 10 phases of the first ChaLearn AutoML challenge. Furthermore, a comprehensive analysis of over 100 diverse datasets, while taking into account time and computational resource constraints, demonstrated that Auto-sklearn outperformed the previous state of the art in AutoML [22]. More details about Auto-sklearn can be found elsewhere [21, 22; https://automl.github.io/auto-sklearn/master/api.html, accessed at 2019-12-10].

Naïve Bayes classifier is a basic ML method that can predict class membership probabilities, such as the probability that a given MDD patient is still depressed after 2 years, with the underlying assumption that the effect of an attribute value on a given class is independent of the values of the other attributes. It aims to simplify the computation involved and, in this sense, is considered naïve [23]. For the present study, we used the Gaussian Naïve Bayes Classifier provided in the `scikit-learn` package with the var_smoothing hyper-parameter. According to the `scikit-learn` manual, by using this implementation a researcher need not choose the probability cut off. Several hyper-parameter settings were tried in the preliminary analysis, resulting in no significant differences. Therefore, the default hyper-parameter setting was used (i.e., setting the value of var_smoothing to `1e-9`). More details about the `scikit-learn` can be found elsewhere (https://scikit-learn.org/stable/modules/generated/sklearn.naive_Bayes.GaussianNB.html #sklearn.naive_Bayes.GaussianNB, accessed at 2019-12-10).

Logistic regression is a classification method used for binary or multinomial outcome variables. Multinomial logistic regression is a classification method that generalizes logistic regression to multiclass problems [31]. We used the `R` package `nnet` [R Foundation for Statistical Computing, Vienna, Austria, 2016. https://www.R-project.org/; 32].

We computed all models by randomly splitting (50:50) the dataset into a training and a test dataset using `Scikit-learn` data split [33]. The training dataset was used to select the best fitting regression model or ML algorithm. For the present study, models were optimized for overall accuracy. Auto-sklearn feature selection and preprocessing were based on the training data. Auto-sklearn selected "multinomial_nb" as its classifier for the binary outcome analysis and "random forest" for the multinomial outcome analyzes. Subsequently, we tested and compared the accuracy of how well these models/algorithms predicted outcomes in the test data with a 95% CI (i.e., percentage of correctly predicted individuals). We also tested and compared their balanced accuracy, sensitivity, specificity, positive predictive value, and negative predictive value. For the multinominal outcomes, this was computed using a one-versus-all approach. For each model, we tested the significance of accuracy related to the no-information rate. The no-information rate contains the accuracy if the model were to choose the most frequent outcome group: healthy, that is, the proportion of correct predictions when all patients are predicted to be healthy. Auto-sklearn and naïve Bayes classifier were implemented using the Python programming language [34]. For logistic regression, `R` was used [R Foundation for Statistical Computing, Vienna, Austria, 2016. https://www.R-project.org/; 32].

7.3 Results

7.3.1 Sociodemographic and clinical characteristics at baseline

Characteristics of the study population are presented in supplementary Table 2. Age at baseline ranged from 18 to 64 years (*M* = 42.2, *SD* = 13.1), and 1,975 (66.5%) participants were women. At baseline, 26.8% of the sample suffered from MDD (*n* = 796), 9.3% of the sample from dysthymia (*n* = 241), and 43.7% from a (comorbid) anxiety disorder (*n* = 1,299), of which social anxiety disorder was the most common (18.6%; *n* = 483). Of the participants in our sample, 46.1% did not meet DSM-IV-TR criteria for a mood or anxiety diagnosis within the preceding 6 months (*n* = 1,368), of whom 54.2% had never been diagnosed with a psychiatric disorder (*n* = 742).

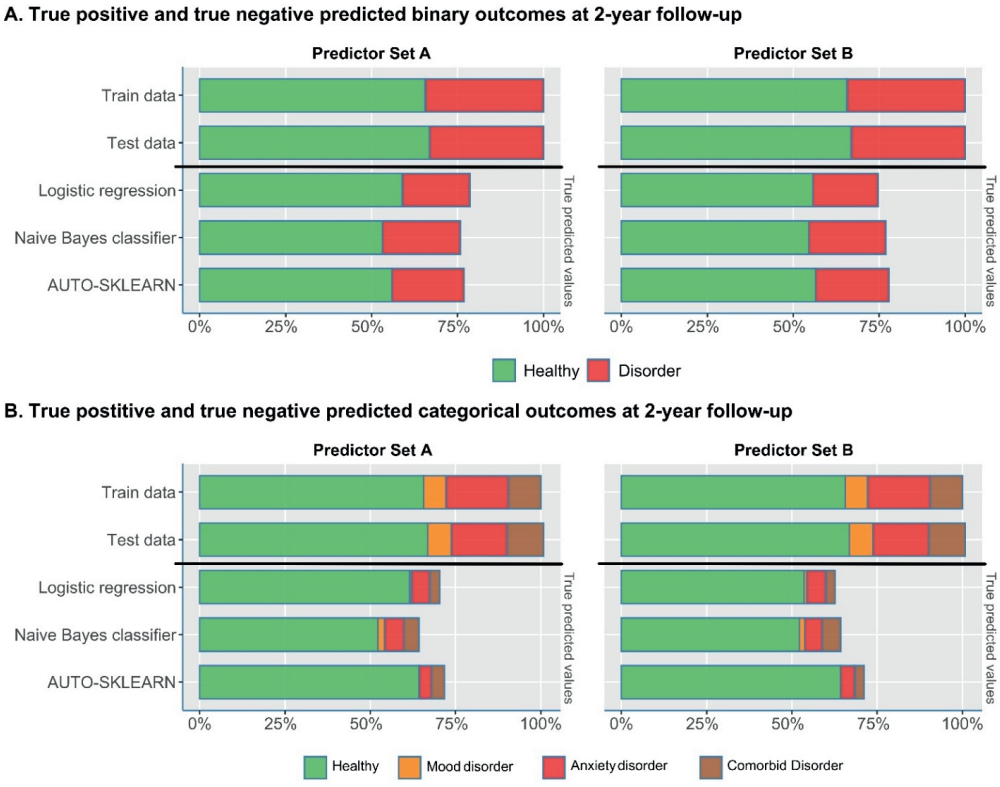


Figure 1. Percentages of train and test dataset values, as well as those correctly predicted at 2-year follow up, using the three data models. All predictor sets included baseline psychiatric diagnoses and demographic variables. Predictor Set A further includes baseline and 1-year follow-up sum scores. Predictor Set B additionally includes baseline and 1-year follow-up individual items.

7.3.2 Prediction of health status as binary outcome

Figures 1 and 2 and supplementary material Figure 1 and Table 3 contain the prediction of health status as a binary outcome (i.e., mentally healthy vs. any anxiety or mood disorder) at the 2-, 4-, 6-, and 9-year follow up using either logistic regression, naïve Bayes classifier, or Auto-sklearn. Figure 1 demonstrates the correctly predicted health status at the 2-year follow up (true negatives and true positives). With optimized overall accuracy, the three methods had different sensitivity and specificity levels. As demonstrated in Figure 2, Auto-sklearn had the highest specificity, with values between .84 and .90, but it had poor sensitivity values (.54–.75), predicting more disorders at the expense of correctly predicting a healthy health status (see also supplementary Table 1). The naïve Bayes classifier had specificity values between .76 and .88 and sensitivity values between .60 and .69. Logistic regression models had the lowest specificity values (.35–.59) but performed better regarding sensitivity values (.82–.93). Together this resulted in balanced accuracy levels ranging from .60–.75, .68–.75, and .63–.74 for Auto-sklearn, naïve Bayes classifier, and logistic regression, respectively.

As further demonstrated in Figure 2, the accuracy values ranged from .75 through .79. Logistic regression, naïve Bayes classifier, and Auto-sklearn were all significantly ($p < .001$) more accurate than the no-information rate (level of accuracy when only predicting a healthy status). Regarding logistic regression, the level of accuracy was significantly higher when only sum scores, and not individual item scores, were included as predictor variables (predictor Set A; acc .79 [95% CI .76–.81]), compared to logistic regression predictor Set B (acc .75 [95% CI .72–.77). The level of accuracy of naïve Bayes classifier and Auto-sklearn did not significantly decrease or improve when individual items were added as predictor variables. At 4-, 6-, and 9-year follow up, accuracy values ranged between .73–.78, .71–.77, and .76–.79 for logistic regression, naïve Bayes classifier, and Auto-sklearn, respectively. Of 16 tests per method (of which eight are presented in Figure 2 and eight in supplementary Table 3), Auto-sklearn had significantly higher accuracy levels than the no-information rate for all tests, compared to eight out of 16 for naïve Bayes classifier and eight out of 16 for logistic regression. Auto-sklearn thus performed adequately within each of the different datasets four different datasets.

	Predictor sets	Sensitivity	Specificity	PPV	NPV	Balanced accuracy	Accuracy
Outcome at year 2:							
- Logistic regression	A.	0.882	0.592	0.813	0.714	0.702	0.786***
	B.	0.834	0.571	0.796	0.631	0.737	0.747***
- Naive Bayes classifier	A.	0.684	0.795	0.624	0.835	0.740	0.758***
	B.	0.677	0.815	0.646	0.836	0.770	0.747***
- AUTO-SKLEARN	A.	0.633	0.836	0.658	0.821	0.735	0.769***
	B.	0.645	0.845	0.675	0.827	0.745	0.779***
Outcome at year 4:							
- Logistic regression	A.	0.911	0.446	0.805	0.668	0.679	0.779***
	B.	0.824	0.490	0.802	0.527	0.657	0.729
- Naive Bayes classifier	A.	0.647	0.808	0.574	0.851	0.728	0.762***
	B.	0.662	0.789	0.556	0.854	0.753	0.753**
- AUTO-SKLEARN	A.	0.414	0.921	0.676	0.797	0.667	0.776***
	B.	0.539	0.885	0.651	0.828	0.712	0.786***
Outcome at year 6:							
- Logistic regression	A.	0.919	0.361	0.808	0.605	0.640	0.777**
	B.	0.858	0.410	0.809	0.498	0.634	0.744
- Naive Bayes classifier	A.	0.639	0.808	0.533	0.867	0.724	0.765
	B.	0.604	0.799	0.507	0.855	0.702	0.749
- AUTO-SKLEARN	A.	0.375	0.930	0.647	0.813	0.652	0.788***
	B.	0.444	0.050	0.615	0.826	0.675	0.787***
Outcome at year 9:							
- Logistic regression	A.	0.928	0.345	0.803	0.626	0.637	0.778**
	B.	0.928	0.472	0.835	0.696	0.700	0.811***
- Naive Bayes classifier	A.	0.633	0.767	0.486	0.857	0.700	0.732
	B.	0.596	0.754	0.457	0.843	0.675	0.713
- AUTO-SKLEARN	A.	0.255	0.949	0.636	0.786	0.602	0.770*
	B.	0.626	0.943	0.614	0.786	0.602	0.767*

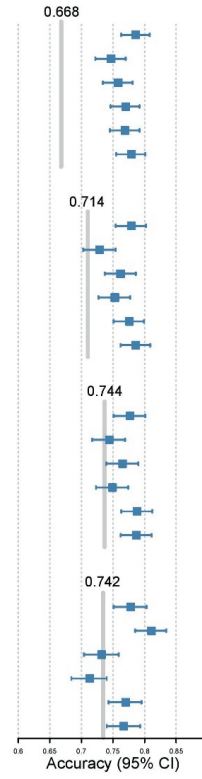


Figure 2. Predicting health status (binary outcome) at 2-, 4-, 6-, and 9-year follow up. All predictor sets included baseline psychiatric diagnoses and demographic variables. Predictor Set A further includes baseline and 1-year follow-up sum scores. Predictor Set B additionally includes baseline and 1-year follow-up individual items. The grey vertical line denotes as the no information rate for year 2-, 4-, 6-, and 9-year outcomes, respectively. Accuracy values were compared to the no-information rate by using a one way ANOVA test of which the *p* values are as follows:

* *p* value < .05

** *p* value < .01

*** *p* value < .001

7.3.3 Prediction of health status as categorical outcome

The results of predicting health status as a categorical outcome (i.e., healthy, mood disorder, anxiety disorder, or comorbid mood- and anxiety disorder) at the 2-, 4-, 6-, and 9-year follow up using either Auto-sklearn, naïve Bayes classifier, or logistic regression are shown in Figures 1, 3, and 4 and in the supplementary material Figure 1 and Tables 4 and 5. Figure 1 demonstrates the correctly predicted health status at 2-year follow up (true positives and true negatives). When the models were optimized for overall accuracy, their performance for predicting the disorder categories were low. When predicting with logistic regression, balanced accuracy values were .53 for mood disorders, .62 for anxiety disorders, and .61 for comorbidity. When predicting with Auto-sklearn, balanced accuracy values were .50 for mood disorders, .60 for anxiety disorders, and .61 for comorbidity. Comparatively, these figures were .70 and .66 when predicting a healthy health status with logistic regression and Auto-sklearn, respectively (see figure 3 outcome year 2). Mood disorder (n = 91 cases in the test data set) was predicted the least often, resulting in sensitivity values ranging from .00–.32 and specificity values ranging from .89–1.00. Further inspection of Figure 1 in the supplementary material demonstrates that both logistic regression and Auto-sklearn mostly predicted a healthy health status instead of mood disorders (n = 55 and n = 68, respectively).

As further demonstrated in Figures 3 and 4, the accuracy values when predicting health status at 2-year follow up ranged from .63 to .72. Both logistic regression (acc .70 [95% CI .68–.73]; p = .003) and Auto-sklearn (acc.72 [95% CI .69–.74]; p < .001) were significantly more accurate than the no-information rate, when predicting health status with sum scores at 2-year follow-up (see Figure 3), but only Auto-sklearn was significantly more accurate than the no-information rate when also individual item scores were included (acc .71 [95% CI .69–.74]; p < .001; see Figure 4). Again, the level of accuracy of logistic regression was significantly lower when individual item scores were included as predictor variables (predictor Set B; acc .63 [95% CI .60–.65]; p = >.99), compared to only sum scores (predictor Set A; acc .70 [95% CI .68–.73]; p = .003) when predicting health status at 2-year follow up. Auto-sklearn achieved demonstrated similar predictive performance when using sum scores as well as individual item scores (see Tables 4 and 5 in the supplementary material). Naïve Bayes classifier did not achieve levels of accuracy above the no-information rate. Achieving significantly accurate predictions became more difficult at later follow-ups. None of the models achieved accuracy levels that exceeded the no-information rate when predicting health status at 4-, 6-, and 9-years follow up.

	Health status	Sensitivity	Specificity	PPV	NPP	Balanced accuracy	Accuracy
Outcome at year 2:							
- Logistic regression	Healthy	0.922	0.480	0.781	0.753	0.701	0.703**
	Mood disorder	0.074	0.988	0.300	0.941	0.531	
	Anxiety disorder	0.333	0.909	0.418	0.874	0.621	
	Comorbidity	0.270	0.959	0.435	0.918	0.614	
- Naive Bayes classifier	Healthy	0.782	0.705	0.842	0.617	0.744	0.643
	Mood disorder	0.321	0.891	0.164	0.952	0.606	
	Anxiety disorder	0.343	0.881	0.361	0.872	0.612	
	Comorbidity	0.423	0.936	0.439	0.932	0.680	
- AUTO-SKLEARN	Healthy	0.963	0.364	0.753	0.831	0.656	0.718***
	Mood disorder	0.000	1.000	0.000	0.938	0.500	
	Anxiety disorder	0.225	0.959	0.516	0.863	0.603	
	Comorbidity	0.358	0.960	0.510	0.927	0.611	
Outcome at year 4:							
- Logistic regression	Healthy	0.944	0.309	0.774	0.688	0.627	0.718
	Mood disorder	0.035	0.986	0.111	0.954	0.511	
	Anxiety disorder	0.173	0.946	0.360	0.867	0.560	
	Comorbidity	0.178	0.972	0.380	0.924	0.575	
- Naive Bayes classifier	Healthy	0.787	0.671	0.857	0.557	0.729	0.668
	Mood disorder	0.193	0.936	0.131	0.959	0.565	
	Anxiety disorder	0.408	0.866	0.348	0.893	0.637	
	Comorbidity	0.402	0.931	0.361	0.941	0.666	
- AUTO-SKLEARN	Healthy	0.992	0.146	0.744	0.877	0.574	0.732
	Mood disorder	0.000	1.000	0.000	0.953	0.499	
	Anxiety disorder	0.089	0.983	0.485	0.860	0.531	
	Comorbidity	0.112	0.989	0.500	0.919	0.559	
Outcome at year 6:							
- Logistic regression	Healthy	0.957	0.240	0.786	0.657	0.957	0.745
	Mood disorder	0.053	0.991	0.231	0.952	0.053	
	Anxiety disorder	0.168	0.963	0.419	0.879	0.168	
	Comorbidity	0.092	0.978	0.233	0.937	0.092	
- Naive Bayes classifier	Healthy	0.795	0.701	0.886	0.540	0.748	0.685
	Mood disorder	0.228	0.950	0.194	0.959	0.589	
	Anxiety disorder	0.426	0.874	0.349	0.905	0.650	
	Comorbidity	0.342	0.913	0.220	0.950	0.627	
- AUTO-SKLEARN	Healthy	0.955	0.278	0.794	0.678	0.619	0.748
	Mood disorder	0.018	0.994	0.143	0.950	0.515	
	Anxiety disorder	0.187	0.950	0.372	0.880	0.567	
	Comorbidity	0.158	0.980	0.364	0.942	0.607	
Outcome at year 9:							
- Logistic regression	Healthy	0.954	0.217	0.778	0.624	0.586	0.737
	Mood disorder	0.019	0.989	0.083	0.950	0.504	
	Anxiety disorder	0.127	0.966	0.388	0.867	0.546	
	Comorbidity	0.154	0.977	0.313	0.945	0.566	
- Naive Bayes classifier	Healthy	0.753	0.648	0.860	0.477	0.700	0.633
	Mood disorder	0.192	0.942	0.149	0.957	0.567	
	Anxiety disorder	0.280	0.893	0.307	0.880	0.586	
	Comorbidity	0.385	0.862	0.157	0.954	0.623	
- AUTO-SKLEARN	Healthy	0.999	0.019	0.745	0.833	0.532	0.745
	Mood disorder	0.000	1.000	0.000	0.950	0.500	
	Anxiety disorder	0.020	0.999	0.750	0.857	0.521	
	Comorbidity	0.015	0.999	0.500	0.938	0.535	

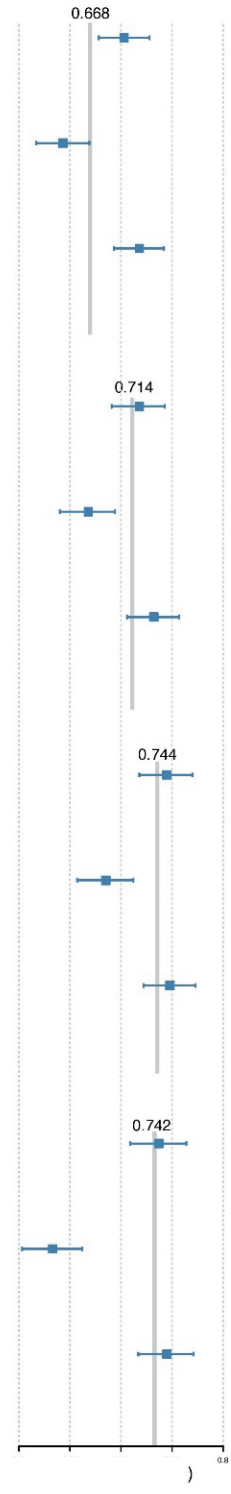


Figure 3. Predicting health status (multinominal outcome) at 2-, 4-, 6-, and 9-year follow up with baseline and 1-year sum scores (predictor Set A). All predictor sets included baseline psychiatric diagnoses and demographic variables. Predictor Set A further includes baseline and 1-year follow-up sum scores. Predictor Set B additionally includes baseline and 1-year follow-up individual items. PPV denotes as positive predictive value. NPV denotes as negative predictive value. The grey vertical line denotes as the no information for year 2-, 4-, 6-, and 9-year outcome, respectively. Accuracy values were compared to the no-information rate by using a one way ANOVA test of which the *p* values are as follows:

** *p* value < .01

*** *p* value < .001

	Health status	Sensitivity	Specificity	PPV	NPP	Balanced accuracy	Accuracy
Outcome at year 2:							
- Logistic regression	Healthy	0.802	0.596	0.800	0.599	0.699	0.627
	Mood disorder	0.136	0.926	0.109	0.942	0.531	
	Anxiety disorder	0.343	0.882	0.363	0.872	0.612	
	Comorbidity	0.255	0.921	0.276	0.913	0.588	
- Naive Bayes classifier	Healthy	0.781	0.717	0.847	0.619	0.749	0.643
	Mood disorder	0.259	0.915	0.168	0.949	0.587	
	Anxiety disorder	0.315	0.894	0.368	0.869	0.604	
	Comorbidity	0.511	0.895	0.365	0.939	0.703	
- AUTO-SKLEARN	Healthy	0.962	0.350	0.749	0.821	0.664	0.712***
	Mood disorder	0.000	1.000	0.000	0.938	0.500	
	Anxiety disorder	0.258	0.948	0.495	0.867	0.592	
	Comorbidity	0.255	0.967	0.479	0.917	0.659	
Outcome at year 4:							
- Logistic regression	Healthy	0.824	0.502	0.805	0.533	0.663	0.646
	Mood disorder	0.105	0.935	0.075	0.955	0.520	
	Anxiety disorder	0.207	0.903	0.272	0.867	0.555	
	Comorbidity	0.243	0.926	0.243	0.926	0.584	
- Naive Bayes classifier	Healthy	0.767	0.685	0.859	0.540	0.726	0.645
	Mood disorder	0.140	0.928	0.089	0.956	0.534	
	Anxiety disorder	0.313	0.880	0.313	0.880	0.596	
	Comorbidity	0.495	0.897	0.319	0.948	0.696	
- AUTO-SKLEARN	Healthy	0.987	0.160	0.746	0.833	0.569	0.729
	Mood disorder	0.000	0.998	0.000	0.952	0.500	
	Anxiety disorder	0.084	0.978	0.405	0.859	0.536	
	Comorbidity	0.131	0.988	0.519	0.921	0.551	
Outcome at year 6:							
- Logistic regression	Healthy	0.806	0.448	0.81	0.442	0.806	0.657
	Mood disorder	0.211	0.927	0.133	0.957	0.211	
	Anxiety disorder	0.226	0.921	0.313	0.882	0.226	
	Comorbidity	0.224	0.931	0.189	0.943	0.224	
- Naive Bayes classifier	Healthy	0.780	0.639	0.863	0.499	0.709	0.662
	Mood disorder	0.246	0.941	0.182	0.959	0.593	
	Anxiety disorder	0.271	0.897	0.296	0.885	0.584	
	Comorbidity	0.474	0.892	0.240	0.959	0.683	
- AUTO-SKLEARN	Healthy	0.957	0.281	0.795	0.692	0.616	0.754
	Mood disorder	0.035	0.994	0.250	0.951	0.506	
	Anxiety disorder	0.174	0.959	0.403	0.879	0.568	
	Comorbidity	0.237	0.977	0.429	0.947	0.569	
Outcome at year 9:							
- Logistic regression	Healthy	0.773	0.453	0.803	0.410	0.613	0.629
	Mood disorder	0.135	0.945	0.115	0.954	0.540	
	Anxiety disorder	0.287	0.882	0.293	0.880	0.585	
	Comorbidity	0.108	0.918	0.080	0.939	0.513	
- Naive Bayes classifier	Healthy	0.723	0.648	0.855	0.448	0.685	0.624
	Mood disorder	0.269	0.925	0.159	0.960	0.597	
	Anxiety disorder	0.313	0.882	0.311	0.883	0.598	
	Comorbidity	0.462	0.879	0.204	0.961	0.670	
- AUTO-SKLEARN	Healthy	0.993	0.071	0.755	0.792	0.509	0.749
	Mood disorder	0.000	1.000	0.000	0.950	0.500	
	Anxiety disorder	0.047	0.994	0.583	0.860	0.509	
	Comorbidity	0.077	0.993	0.417	0.941	0.507	

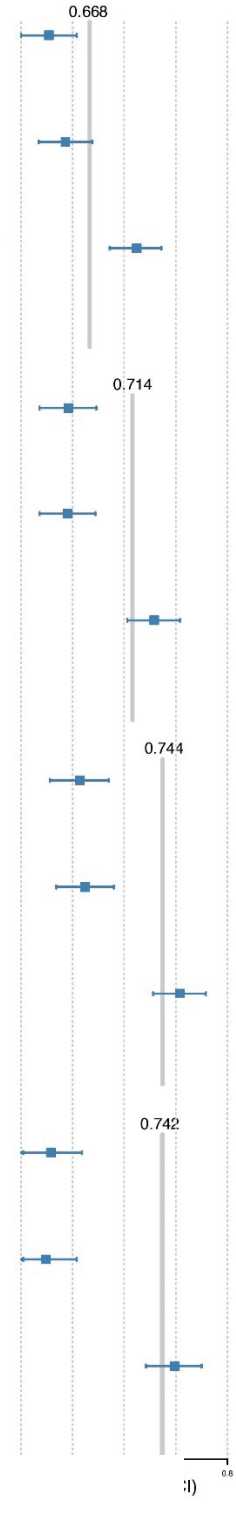


Figure 4. Predicting health status (multinominal outcome) at 2-, 4-, 6-, and 9-year follow up with baseline and 1-year sum scores and individual item-scores (predictor Set B). All predictor sets included baseline psychiatric diagnoses and demographic variables. Predictor Set B further includes baseline and 1-year follow-up sum scores and individual items. PPV denotes as positive predictive value. NPV denotes as negative predictive value. The grey vertical line denotes as the no information rate for year 2-, 4-, 6-, and 9-year outcome, respectively. Accuracy values were compared to the no-information rate by using a one way ANOVA test of which the p values are as follows: *** p value < .001

7.4 Discussion

Our aim was to assess and compare the predictive performances and clinical usefulness of Auto-sklearn, naïve Bayes classifier, and logistic regression to predict mood and anxiety disorders at follow up. Furthermore, we assessed the effects of different sets of predictors. Although we hypothesized that Auto-sklearn would outperform the two other data models, this could not be concluded unequivocally. In fact, only moderate levels of accuracy were found, with correct prediction percentages of up to 79% and 75% when using either binary or categorical outcomes, respectively. Yet, Auto-sklearn outperformed both logistic regression and naïve Bayes when predictor sets included individual item scores. Categorical outcomes were more difficult to predict than binary outcomes, compared to the no-information rate; in particular, mood disorders could not be distinguished well.

Our results support those of previous ML studies that reported 60% to 82% of correctly predicted mood and anxiety diagnoses when using a broad spectrum of predictor variables [10, 12-14, 35, 36]. One of these studies used a subset of the NESDA dataset that included patients with a depression at baseline and a more extensive set of clinical, behavioral, and biological baseline-only variables in order to predict the course of depression, resulting in accuracy levels of 62–66% [13]. A similar study, within a subset of anxiety patients in NESDA (again using an extensive set of predictors) found an accuracy for predicting anxiety recovery of 62% and a accuracy of predicting recovery of all common mental disorders of 63% [14]. In contrast to these prior studies, we only used data that could be easily collected in clinical practice, including 1-year follow-up data as predictor variables. Despite our dataset not being as rich and diverse, we achieved a higher overall accuracy which was significantly higher than the no-information rate [13, 14]. However, these results cannot be compared easily. Our often higher accuracy values were likely in part due to our inclusion of healthy participants. The predictive performance when predicting the disorder value were similar and the large proportion of the healthy health status outcomes resulted in unbalanced sensitivity and specificity values when models were optimized to maximum overall accuracy. Prior studies lacked thorough comparisons to (logistic) regression models, and thereby failed to address the additional value of ML methods over “traditional” data-modelling methods.

Previous ML studies in the field of psychiatry used a wide variety of ML methods, ranging from regression trees to gradient boosting machines—methods that were included in Auto-sklearn [10, 35]. In line with an earlier study, we found that depending on the predictor set, more complex ML

methods do not necessarily result in higher similar levels of accuracy when predicting future outcomes of mood disorders [36]. Two previous studies found that when optimized on overall level of accuracy, ML methods were about 1–6% more accurate compared to regression analysis and needed fewer predictor variables when predicting the persistence of mood disorders at a 12-week follow up [10, 35]. Although level of accuracy was higher for ML, this difference was not found to be significant in either study [10, 35]. Several studies found that ML was of only limited added value in research (Belsher et al., 2019; Christodoulou et al., 2019; van Mens et al., 2020) and clinical usefulness [19]. Although we did not find any published reviews within the field of psychiatry, within other fields the added value of ML has been notably criticized [e.g., 16, 37, 38]. However, it is possible that ML does outperform traditional methods when more complex (large) datasets are used [7, 8]. More advanced ML methods have the capability to distinguish which variables in large datasets are relevant or irrelevant for prediction, whereas traditional (regression) models rely on the researcher or clinician to select variables of interest to a particular analysis. ML therefore requires less human input. Although regression models sequentially analyze the relationship between variables, ML approaches can iteratively and contemporaneously analyze multiple interacting associations between variables or variable sets. Indeed, ML approaches may potentially be better suited to complex datasets with a large amount of predictors, while limiting the risk of overfitting [12]. These advantages were confirmed by our findings. Auto-sklearn outperformed the other two models when our predictor sets included more variables, that is, they were more complex.

ML, especially when automated, has the potential for use in mental healthcare. Deciding what information to collect from patients and making predictions on the micro and macro level based on that information are important aspects of a clinician's skill set. This includes predictions regarding suicide risk, violence, the efficacy of treatment options, and the prognoses on the course of disorders [2]. The accuracy of these predictions is of vital importance for individual patients. Two major approaches to predict clinical outcomes can be identified: the clinical and the statistical method. The clinical approach refers to an informal and intuitive process in which the clinician combines and integrates patient data. A clinician's experience, interpersonal sensitivity, and theoretical perspective combined with a patient's characteristics and circumstances determine how that clinician recalls, synthesizes, and interprets all these bits of information [2]. With a statistical approach, statistical methods are applied on objectively measured variables in order to make predictions and prognoses based on probabilities [2]. Two meta-analyzes demonstrated that

statistical approaches were more accurate than clinical methods [2, 5]. Our study found that moderate levels of accuracy can be accomplished based on data that can be easily collected in clinical practice, confirming that integrating statistical methods into clinical decision making could provide added benefits. Current mental healthcare is already partly digitalized, and the development of automated digital tools to assist clinicians should be attainable, providing clinicians fast and cheap support in decision making. Automated ML can be developed into such a tool because its automated techniques can match or improve upon expert human performance in certain ML tasks—often in a shorter amount of time [20]. Moreover, Auto-sklearn demonstrated that it can perform even under rigid time and computational resource constraints [21]. Automated ML is already demonstrating its usefulness in healthcare practice [20].

There are several study limitations that need to be discussed. First, despite the marginal differences between DSM-IV-TR and DSM-5 criteria for mood and anxiety disorders, the diagnostic classifications used in this study were slightly outdated but were chosen to be kept constant during the follow-up waves [39]. Despite our relatively large sample size, our analyzes could not be carried out for each diagnosis separately (e.g., dysthymia, panic disorder, etc.) because the samples would have become too small. Second, in contrast with other studies, we did not replicate our findings with an independent dataset [10, 36]. Although we made use of a training and testing dataset, it is possible that the results from the ML methods and regression analyzes differed in generalizability to other datasets, which could not be assessed with our current study design. Third, NESDA is an observational cohort study, and different types of pharmacological and psychotherapeutic treatment were not taken into account as predictor variables. Fourth, we included both healthy participants and patients, testing concomitantly the prediction of the course and onset of depression and anxiety. The proportion of healthy controls may have influenced the predictive models because their homeostatic responses to internal or external stimuli do not represent that of psychopathologic disorders [40]. The large proportion of the healthy health status outcomes resulted in unbalanced sensitivity and specificity values when models were optimized to maximum overall accuracy. Fifth, differentiating depression, anxiety, and comorbid disorders as multinomial variables was especially poor and may have been unrealistic because anxiety disorders and depression have overlapping risk factors and high levels of (subclinical) comorbidity [41, 42]. Sixth, ML may have more added value when the dataset is more complex, such as imaging or genetic data [7, 8, 12]. Although our data was easy to collect in clinical practice, it may have lacked the complexity that is needed for ML methods to excel. Finally, because of its automated features, Auto-sklearn

acts like a black box, which made it difficult for us to examine which individual features were most predictive. Nevertheless, significant levels of accuracy were achieved when predictor sets included sociodemographic, baseline diagnoses, and self-reported sum scores, which did not significantly improve when variables were added, suggesting that these were the most important predictor variables.

In conclusion, we found that moderately high levels of accuracy could be achieved when predicting dichotomous outcomes with easy-to-collect data. Auto-sklearn did not achieve the highest level of accuracy in every set of predictors, compared to traditional logistic regression and a naïve Bayes classifier. However, it was most consistent regardless of the set of predictor variables, and it outperformed the other models when the predictor sets were more complex (i.e., individual item scores). In time, clinical practice may benefit from integrating next generation automated ML methods into clinical decision making.

Acknowledgement

The infrastructure for the NESDA study (www.nesda.nl) is funded through the Geestkracht program of the Netherlands Organization for Health Research and Development (ZonMw, grant number 10-000-1002) and financial contributions by participating universities and mental healthcare organizations (Amsterdam University Medical Centers (location VUmc), GGZ inGeest, Leiden University Medical Center, Leiden University, GGZ Rivierduinen, University Medical Center Groningen, University of Groningen, Lentis, GGZ Friesland, GGZ Drenthe, Rob Giel Onderzoekscentrum).

References

1. McGorry, P.D., *Risk syndromes, clinical staging and DSM V: new diagnostic infrastructure for early intervention in psychiatry.* Schizophrenia research, 2010. **120**(1): p. 49-53.
2. Ægisdóttir, S., et al., *The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction.* The Counseling Psychologist, 2006. **34**(3): p. 341-382.
3. Odeh, M.S., R.A. Zeiss, and M.T. Huss, *Cues they use: Clinicians' endorsement of risk cues in predictions of dangerousness.* Behavioral sciences & the law, 2006. **24**(2): p. 147-156.
4. Carlier, I.V., et al., *Routine outcome monitoring and feedback on physical or mental health status: evidence and theory.* Journal of Evaluation in Clinical Practice, 2012. **18**(1): p. 104-110.
5. Grove, W.M., et al., *Clinical versus mechanical prediction: a meta-analysis.* Psychological assessment, 2000. **12**(1): p. 19.
6. Johnson, A.E., et al., *Machine learning and decision support in critical care.* Proceedings of the IEEE. Institute of Electrical and Electronics Engineers, 2016. **104**(2): p. 444.
7. Iniesta, R., D. Stahl, and P. McGuffin, *Machine learning, statistical learning and the future of biological research in psychiatry.* Psychological medicine, 2016. **46**(12): p. 2455-2465.
8. Wang, Y., L. Kung, and T.A. Byrd, *Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations.* Technological Forecasting and Social Change, 2018. **126**: p. 3-13.
9. Hahn, T., A. Nierenberg, and S. Whitfield-Gabrieli, *Predictive analytics in mental health: applications, guidelines, challenges and perspectives.* Molecular psychiatry, 2016. **22**(1): p. 37.
10. Chekroud, A.M., et al., *Cross-trial prediction of treatment outcome in depression: a machine learning approach.* The Lancet Psychiatry, 2016. **3**(3): p. 243-250.
11. Hastie, T., R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction, Springer Series in Statistics*. 2009, Springer New York.
12. Lee, Y., et al., *Applications of machine learning algorithms to predict therapeutic outcomes in depression: A meta-analysis and systematic review.* Journal of affective disorders, 2018.
13. Dinga, R., et al., *Predicting the naturalistic course of depression from a wide range of clinical, psychological, and biological data: a machine learning approach.* Translational psychiatry, 2018. **8**(1): p. 241.
14. Bokma, W.A., et al., *Predicting the naturalistic course in anxiety disorders using clinical and biological markers: a machine learning approach.* Psychological Medicine, 2020: p. 1-11.
15. van Mens, K., et al., *Predicting future suicidal behaviour in young adults, with different machine learning techniques: a population-based longitudinal study.* Journal of affective disorders, 2020.
16. Christodoulou, E., et al., *A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models.* Journal of clinical epidemiology, 2019.
17. Kessler, R.C., et al., *Developing a practical suicide risk prediction model for targeting high-risk patients in the Veterans Health Administration.* International journal of methods in psychiatric research, 2017. **26**(3): p. e1575.
18. Belsher, B.E., et al., *Prediction models for suicide attempts and deaths: a systematic review and simulation.* JAMA psychiatry, 2019. **76**(6): p. 642-651.
19. Tran, B.X., et al., *The current research landscape on the artificial intelligence application in the management of depressive disorders: A bibliometric analysis.* International journal of environmental research and public health, 2019. **16**(12): p. 2150.
20. Waring, J., C. Lindvall, and R. Umeton, *Automated machine learning: Review of the state-of-the-art and opportunities for healthcare.* Artificial Intelligence in Medicine, 2020: p. 101822.
21. Feurer, M., et al. *Efficient and robust automated machine learning*. in *Advances in Neural Information Processing Systems*. 2015.
22. Feurer, M., et al., *Auto-sklearn: Efficient and Robust Automated Machine Learning*, in *Automated Machine Learning*. 2019, Springer. p. 113-134.
23. Jayant, A. and a.O.R.M.C. Safari, *Data Science and Machine Learning Series: Naive Bayes Classifier Advanced Concepts*. 2020: Technics Publications.
24. Penninx, B.W., et al., *The Netherlands Study of Depression and Anxiety (NESDA): rationale, objectives and methods.* Int J Methods Psychiatr Res, 2008. **17**(3): p. 121-40.
25. Marks, I.M. and A.M. Mathews, *Brief standard self-rating for phobic patients.* Behaviour research and therapy, 1979. **17**(3): p. 263-267.
26. Beck, A.T., et al., *An inventory for measuring clinical anxiety: psychometric properties.* Journal of consulting and clinical psychology, 1988. **56**(6): p. 893.

27. Rush, A.J., et al., *The Inventory of Depressive Symptomatology (IDS): psychometric properties.* Psychol Med, 1996. **26**(3): p. 477-86.
28. Wittchen, H.-U., *Reliability and validity studies of the WHO-Composite International Diagnostic Interview (CIDI): A critical review.* Journal of Psychiatric Research, 1994. **28**(1): p. 57-84.
29. Haro, J.M., et al., *Concordance of the Composite International Diagnostic Interview Version 3.0 (CIDI 3.0) with standardized clinical assessments in the WHO World Mental Health surveys.* International journal of methods in psychiatric research, 2006. **15**(4): p. 167-180.
30. Kessler, R.C., et al., *National comorbidity survey replication adolescent supplement (NCS-A): III. Concordance of DSM-IV/CIDI diagnoses with clinical reassessments.* Journal of the American Academy of Child & Adolescent Psychiatry, 2009. **48**(4): p. 386-399.
31. Menard, S., *Applied logistic regression analysis*. Vol. 106. 2002: Sage.
32. Ripley, B., W. Venables, and M.B. Ripley, *Package 'nnet'.* R package version, 2016: p. 7-3.
33. Pedregosa, F., et al., *Scikit-learn: Machine learning in Python.* the Journal of machine Learning research, 2011. **12**: p. 2825-2830.
34. Rossum, G.v., *Python tutorial, Technical Report CS-R9526*. 1995, The Netherlands, Amsterdam: Centrum voor Wiskunde en Informatica (CWI).
35. Kessler, R.C., et al., *Testing a machine-learning algorithm to predict the persistence and severity of major depressive disorder from baseline self-reports.* Molecular psychiatry, 2016. **21**(10): p. 1366.
36. Nie, Z., et al., *Predictive modeling of treatment resistant depression using data from STAR*D and an independent clinical study.* PLOS ONE, 2018. **13**(6): p. e0197268.
37. Frizzell, J.D., et al., *Prediction of 30-day all-cause readmissions in patients hospitalized for heart failure: comparison of machine learning and other statistical approaches.* JAMA cardiology, 2017. **2**(2): p. 204-209.
38. Desai, R.J., et al., *Comparison of Machine Learning Methods With Traditional Models for Use of Administrative Claims With Electronic Medical Records to Predict Heart Failure Outcomes.* JAMA Network Open, 2020. **3**(1): p. e1918962-e1918962.
39. Regier, D.A., E.A. Kuhl, and D.J. Kupfer, *The DSM-5: Classification and criteria changes.* World psychiatry, 2013. **12**(2): p. 92-98.
40. Regier, D.A., et al., *Limitations of diagnostic criteria and assessment instruments for mental disorders: implications for research and policy.* Archives of general psychiatry, 1998. **55**(2): p. 109-115.
41. Jacobson, N.C. and M.G. Newman, *Anxiety and depression as bidirectional risk factors for one another: A meta-analysis of longitudinal studies.* Psychological bulletin, 2017. **143**(11): p. 1155.
42. Shorter, E. and P. Tyrer, *Separation of anxiety and depressive disorders: blind alley in psychopharmacology and classification of disease.* BMJ, 2003. **327**(7407): p. 158-160.

Supplementary Material

Measures

Composite International Diagnostic Interview

The Composite International Diagnostic Interview (CIDI WHO, version 2.1) was used to assess the presence of depressive and anxiety disorders according to the DSM-IV-TR. The CIDI is used worldwide and has been demonstrated to have high interrater reliability high test–retest reliability [1, 2] and high validity for depressive and anxiety disorders [3, 4]. Trained clinical research staff conducted the interviews [5].

Fear Questionnaire

The 15-item Fear Questionnaire (FQ) is a self-report instrument that assesses the level of avoidance in relation to common phobias, including social phobia (Items 2, 6, 8,10, 13), agoraphobia (Items 4, 5, 7, 11, 14), and hematophobia/traumatophobia [Items 1, 3, 9, 12, 15; 6]. It consists of 15 equally weighted items, rated on a 9-point scale, ranging from 0 (*Would not avoid it*) to 8 (*Always avoid it*). The sum score ranges from 0 to 120. The psychometric properties of the FQ have been researched in multiple studies among both nonclinical populations and patients with an anxiety disorder [7-9].

Beck's Anxiety Inventory

The 21-item Beck's Anxiety Inventory (BAI) is a self-report instrument that assesses the overall severity of anxiety.[10] The items consist of 21 anxiety symptoms, including physical symptoms (e.g., "Heart pounding/racing") and psychological symptoms (e.g., "Fear of the worst happening"). It consists of equally weighted items, rated on a 4-point scale, ranging from 0 (*not at all*) to 3 (*severely, I could barely stand it*). The BAI is scored by adding the ratings for all 21 symptoms to obtain a total score that can range from 0 to 63. The reliability and validity of the BAI are well established [10, 11].

Inventory of Depressive Symptomatology

The 30-item Inventory of Depressive Symptomatology (IDS-SR) was used to assess the severity of depression [12, 13]. The IDS-SR scale includes all symptoms of depression, including melancholic, atypical, and anxious symptoms. Moreover, several additional symptoms have been added, such as sympathetic arousal, pessimism, and interest in sex. It consists of 30 equally weighted items, rated on a 4-point scale (0–3). The IDS-SR is scored by adding the ratings of the 30 symptoms to obtain a total score that can range from 0 to 88. Items 11 and 12 ("increased/decreased appetite) and Items 13 and 14 (weight gain/weight loss) contain opposite features, so we combined each of them into

two ordinal items with both severe increase or decrease at Scale 3, yielding 28 items for the current analyzes [13].

References

1. Wittchen, H.-U., et al., *Cross-cultural feasibility, reliability and sources of variance of the Composite International Diagnostic Interview (CIDI).* The British Journal of Psychiatry, 1991. **159**(5): p. 645-653.
2. Wacker, H., et al., *Using the CIDI-C in the general population. Psychiatry: A world perspective. Edited by: Stefanis CN, Rabavilas AD, Soldatos CR. 2006*. Amsterdam: Elsevier Science Publishers.
3. Wittchen, H.-U., *Reliability and validity studies of the WHO-Composite International Diagnostic Interview (CIDI): A critical review.* Journal of Psychiatric Research, 1994. **28**(1): p. 57-84.
4. Wittchen, H.-U., et al., *Recall and dating of psychiatric symptoms: test-retest reliability of time-related symptom questions in a standardized psychiatric interview.* Archives of General Psychiatry, 1989. **46**(5): p. 437-443.
5. Penninx, B.W., et al., *The Netherlands Study of Depression and Anxiety (NESDA): rationale, objectives and methods.* Int J Methods Psychiatr Res, 2008. **17**(3): p. 121-40.
6. Marks, I.M. and A.M. Mathews, *Brief standard self-rating for phobic patients.* Behaviour research and therapy, 1979. **17**(3): p. 263-267.
7. Gillis, M.M., D.A. Haaga, and G.T. Ford, *Normative values for the Beck Anxiety Inventory, Fear Questionnaire, Penn State Worry Questionnaire, and Social Phobia and Anxiety Inventory.* Psychological Assessment, 1995. **7**(4): p. 450.
8. Oei, T.P., A. Moylan, and L. Evans, *Validity and clinical utility of the Fear Questionnaire for anxiety-disorder patients.* Psychological Assessment: A Journal of Consulting and Clinical Psychology, 1991. **3**(3): p. 391.
9. Mavissakalian, M., *The Fear Questionnaire: a validity study.* Behaviour Research and Therapy, 1986. **24**(1): p. 83-85.
10. Beck, A.T., et al., *An inventory for measuring clinical anxiety: psychometric properties.* Journal of consulting and clinical psychology, 1988. **56**(6): p. 893.
11. Steer, R.A., et al., *Structure of the computer-assisted Beck Anxiety Inventory with psychiatric inpatients.* Journal of personality assessment, 1993. **60**(3): p. 532-542.
12. Trivedi, M.H., et al., *The Inventory of Depressive Symptomatology, Clinician Rating (IDS-C) and Self-Report (IDS-SR), and the Quick Inventory of Depressive Symptomatology, Clinician Rating (QIDS-C) and Self-Report (QIDS-SR) in public sector patients with mood disorders: a psychometric evaluation.* Psychological Medicine, 2004. **34**(1): p. 73-82.
13. Rush, A.J., et al., *The Inventory of Depressive Symptomatology (IDS): Psychometric properties.* Psychological Medicine, 1996. **26**(3): p. 477-486.

Supplementary Table 1. The variables that were part of the four predictor sets (A through D) that were used in the analyzes with the 3 different data models

	Predictor set A	Predictor set B	Predictor set C	Predictor set D
1.	MDD (yes/no)	MDD (yes/no)	MDD (yes/no)	MDD (yes/no)
2.	Dysthymia (yes/no)	Dysthymia (yes/no)	Dysthymia (yes/no)	Dysthymia (yes/no)
3.	Minor depression (yes/no)	Minor depression (yes/no)	Minor depression (yes/no)	Minor depression (yes/no)
4.	History of MDD (yes/no)	History of MDD (yes/no)	History of MDD (yes/no)	History of MDD (yes/no)
5.	Social-phobia (yes/no)	Social-phobia (yes/no)	Social-phobia (yes/no)	Social-phobia (yes/no)
6.	Panic disorder with agoraphobia (yes/no)	Panic disorder with agoraphobia (yes/no)	Panic disorder with agoraphobia (yes/no)	Panic disorder with agoraphobia (yes/no)
7.	Panic disorder without agoraphobia (yes/no)	Panic disorder without agoraphobia (yes/no)	Panic disorder without agoraphobia (yes/no)	Panic disorder without agoraphobia (yes/no)
8.	Agoraphobia (yes/no)	Agoraphobia (yes/no)	Agoraphobia (yes/no)	Agoraphobia (yes/no)
9.	Generalized anxiety disorder (yes/no)	Generalized anxiety disorder (yes/no)	Generalized anxiety disorder (yes/no)	Generalized anxiety disorder (yes/no)
10.	History of anxiety disorder	History of anxiety disorder	History of anxiety disorder	History of anxiety disorder
11.	Gender	Gender	Gender	Gender
12.	Age	Age	Age	Age
13.	Level of education	Level of education	Level of education	Level of education
14.	North European ancestry (yes/no)	North European ancestry (yes/no)	North European ancestry (yes/no)	North European ancestry (yes/no)
15.	Partner status	Partner status	Partner status	Partner status
16.	Work status	Work status	Work status	Work status
17.	FQ sumscore baseline	FQ sumscore baseline	FQ sumscore baseline	FQ sumscore baseline
18.	BAI sumscore baseline	BAI sumscore baseline	BAI sumscore baseline	BAI sumscore baseline
19.	IDS-SR sumscore baseline	IDS-SR sumscore baseline	IDS-SR sumscore baseline	IDS-SR sumscore baseline
20.	FQ sumscore at 1-year follow-up	FQ item 1 through 15		FQ item 1 through 15
21.	BAI sumscore at 1-year follow-up	BAI item 1 through 21		BAI item 1 through 21
22.	IDS-SR sumscore at 1-year follow-up	IDS-SR item 1 through 28		IDS-SR item 1 through 28
23.	Delta FQ sumscore (1-year follow-up – baseline)	FQ sumscore at 1-year follow-up		
24.	Delta BAI sumscore (1-year follow-up – baseline)	BAI sumscore at 1-year follow-up		
25.	Delta IDS-SR sumscore (1-year follow-up – baseline)	IDS-SR sumscore at 1-year follow-up		
26.		Delta FQ sumscore (1-year follow-up – baseline)		
27.		Delta BAI sumscore (1-year follow-up – baseline)		
28.		Delta IDS-SR sumscore (1-year follow-up – baseline)		
29.		FQ item 1 through 15 at 1-year follow-up		
30.		BAI item 1 through 21 at 1-year 1 follow-up		
31.		IDS-SR item 1 through 28 at 1-year follow-up		

32.	Delta FQ item 1 through 15 (1-year follow up – baseline)
33.	Delta BAI item 1 through 21 (1-year follow up – baseline)
34.	Delta IDS-SR item 1 through 28 (1-year follow up – baseline)
35.	Delta IDS-SR item 28 - year 1 follow up[…]Delta BAI item 21 - year 1 follow up[…]

Note. MDD denotes Major depressive disorder. FQ denotes Fear Questionnaire. BAI denotes Beck's anxiety inventory. IDS-SR denotes as inventory of depressive symptomatology.

Supplementary Table 2. Baseline sociodemographic and clinical characteristics of the 2,596 NESDA participants.

	Cohort
Age in years (mean, *SD*)	42.2 (13.1)
Female (%)	65.5
North-European ethnicity (%)	94.8
Education level (%)	
Elementary or lower	38.1
Secondary education	58.0
College or university	3.9
Work status (%)	
Employed	53.4
Self-employed	6.3
Disability	9.1
Sick benefit	5.0
Early retirement	3.4
Unemployed	18.3
Partner status (%)	
Married	38.5
Partner but was not married	30.8
Divorced	7.3
Widowed	1.4
Mood disorder (%)	
Major depressive disorder	26.8
Minor depression	2.8
Dysthymia	9.3
Lifetime depression	66.2
Anxiety disorder (%)	
Panic disorder with agoraphobia	11.9
Panic disorder without agoraphobia	5.2
Agoraphobia without panic	5.1
Generalized anxiety disorder	13.3
Social anxiety disorder	18.6
Lifetime anxiety disorder	59.4
No Disorder (%)	46.1
No lifetime disorder	24.9
Self reports (mean, *SD*)	
Baseline totalscore IDS-SR	21.5 (14.1)
Baseline totalscore FQ	24.8 (19.9)
Baseline totalscore BAI	12.1 (10.7)
Year-1 totalscore IDS-SR	16.9 (12.4)
Year-1 totalscore FQ	20.8 (18.6)
Year-1 totalscore BAI	9.3 (9.2)

Note. *SD* denotes standard deviation. IDS-SR denotes Inventory of Depressive Symptomatology - Self Report. FQ denotes Fear Questionnaire. BAI denotes Beck Anxiety Inventory.

Supplementary Table 3 Predicting mental health status (binary outcome) at 2-, 4-, 6-, and 9-year follow up using baseline data as the independent variables (i.e., predictor set C, and D).

AUTO-SKLEARN	Outcome Year 2	Outcome Year 4	Outcome Year 6	Outcome Year 9
Baseline sum-scores				
accuracy	0.763	0.764	0.781	0.770
95% CI	0.739 - 0.786	0.738 - 0.787	0.756 - 0.805	0.743 - 0.795
p value [acc > NIR)	<0.001	<0.001	0.003	0.020
balanced accuracy	0.706	0.674	0.640	0.585
sensitivity	0.534	0.466	0.351	0.202
specificity	0.878	0.882	0.929	0.967
positive predictive value	0.685	0.613	0.627	0.684
negative predictive value	0.791	0.805	0.807	0.777
Baseline sum-scores and individual items				
accuracy	0.773	0.770	0.769	0.773
95% CI	0.749 - 0.795	0.745 - 0.794	0.743 - 0.793	0.746 - 0.798
p value [acc > NIR)	<0.001	<0.001	0.034	0.012
balanced accuracy	0.713	0.651	0.610	0.625
sensitivity	0.534	0.373	0.285	0.318
specificity	0.892	0.929	0.935	0.931
positive predictive value	0.710	0.677	0.599	0.616
negative predictive value	0.794	0.788	0.792	0.797
Naive Bayes classifier				
Baseline sum-scores				
accuracy	0.755	0.759	0.762	0.730
95% CI	0.732 - 0.778	0.733 - 0.782	0.736 - 0.786	0.702 - 0.757
p value [acc > NIR)	<0.001	<0.001	0.103	0.813
balanced accuracy	0.734	0.720	0.722	0.696
sensitivity	0.673	0.630	0.642	0.625
specificity	0.796	0.810	0.802	0.767
positive predictive value	0.621	0.570	0.527	0.483
negative predictive value	0.830	0.845	0.867	0.855
Baseline sum-scores and individual items				
accuracy	0.761	0.750	0.750	0.711
95% CI	0.737 - 0.784	0.725 - 0.774	0.724 - 0.775	0.682 - 0.739
p value [acc > NIR)	<0.001	0.003	0.355	0.989
balanced accuracy	0.746	0.725	0.714	0.683
sensitivity	0.701	0.668	0.639	0.625
specificity	0.791	0.783	0.788	0.741
positive predictive value	0.625	0.552	0.508	0.456
negative predictive value	0.842	0.855	0.864	0.851
Logistic regression				
Baseline sum-scores				
accuracy	0.766	0.769	0.762	0.769
95% CI	0.741 - 0.789	0.744 - 0.792	0.737 - 0.787	0.742 - 0.794
P value [acc > NIR)	0.000	0.000	0.091	0.625
balanced accuracy	0.713	0.661	0.619	0.625
sensitivity	0.870	0.911	0.912	0.923
specificity	0.557	0.411	0.326	0.326
positive predictive value	0.798	0.795	0.798	0.798
negative predictive value	0.680	0.650	0.560	0.596
Baseline sum-scores and individual items				
accuracy	0.743	0.748	0.748	0.782
95% CI	0.718 - 0.7663	0.722 - 0.772	0.722 - 0.773	0.755 - 0.807
p value [acc > NIR)	0.000	0.005	0.408	0.002
balanced accuracy	0.696	0.649	0.621	0.646
sensitivity	0.835	0.879	0.881	0.926
specificity	0.557	0.420	0.361	0.367
positive predictive value	0.791	0.791	0.801	0.808
negative predictive value	0.627	0.581	0.510	0.632

Note. The *p* value denotes the one-sided ANOVA statistic of accuracy (acc) compared with the No-Information Rate (NIR). NIR was 0.668, 0.714, 0.744, and 0.742 for year 2-, 4-, 6-, and 9-year outcome, respectively.

Supplementary Table 4 Predicting mental health status (categorical outcome) at 2-, 4-, 6-, and 9-year follow up using baseline sum scores as the independent variables (i.e., predictor Set C)

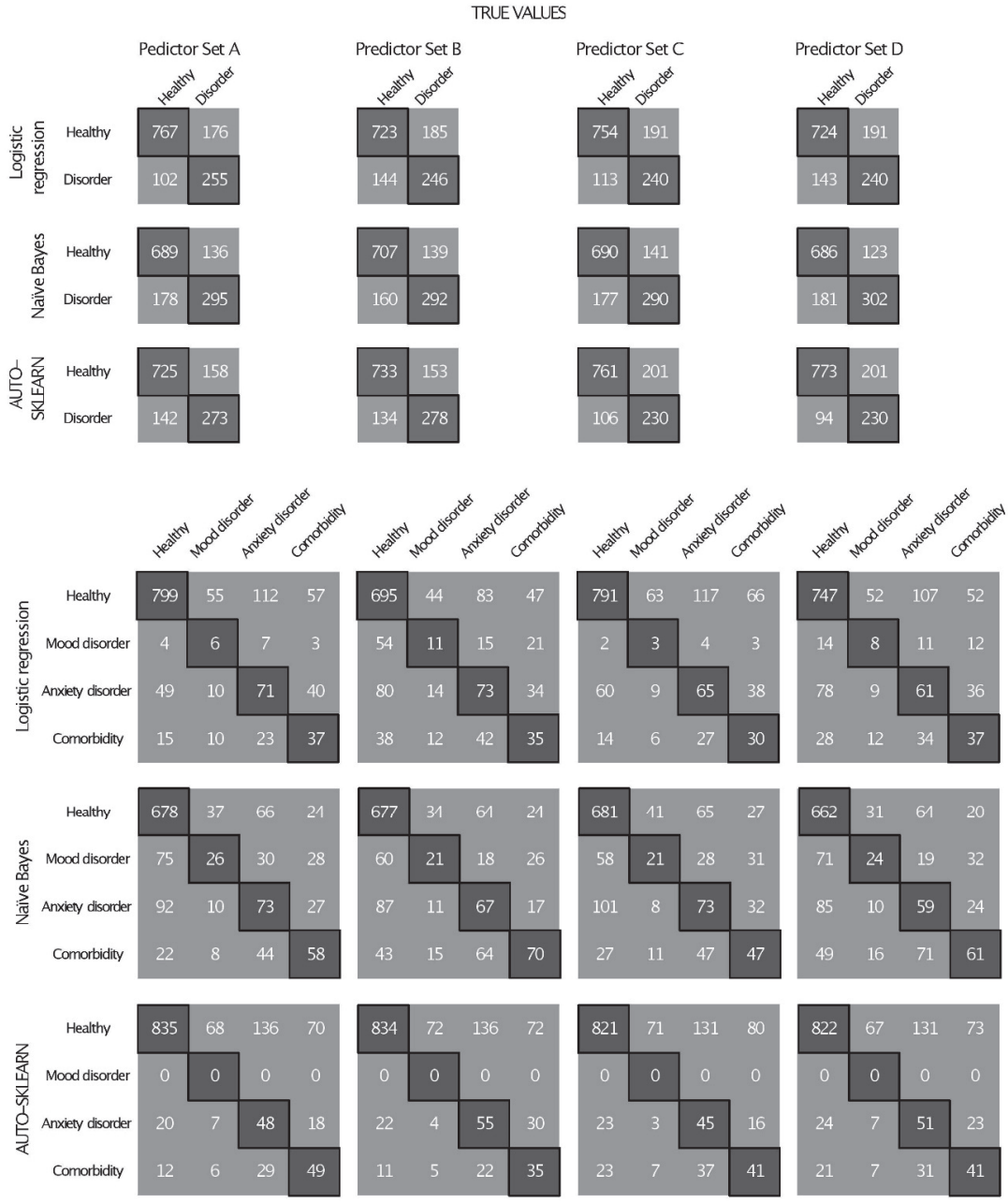
UTO-SKLEARN	Outcome at Year 2				Outcome at Year 4				Outcome at Year 6				Outcome at Year 9			
ealth status	I	II	III	IV	I	II	III	IV	I	II	III	IV	I	II	III	IV
aseline sum scores																
accuracy	0.699				0.729				0.748				0.749			
95% CI	0.673 - 0.724				0.702 - 0.754				0.722 - 0.773				0.721 - 0.775			
p value [acc > NIR)	0.010				0.146				0.408				0.324			
balanced accuracy	0.646	0.500	0.586	0.621	0.569	0.500	0.535	0.570	0.537	0.500	0.527	0.517	0.527	0.500	0.521	0.528
sensitivity	0.947	0.000	0.211	0.299	0.983	0.000	0.084	0.159	0.988	0.000	0.071	0.039	0.995	0.000	0.047	0.062
specificity	0.346	1.000	0.961	0.942	0.155	0.999	0.985	0.982	0.087	1.000	0.984	0.995	0.060	1.000	0.995	0.995
positive predictive value	0.744	0.000	0.517	0.380	0.744	0.000	0.500	0.459	0.759	0.000	0.407	0.375	0.753	0.000	0.636	0.444
negative predictive value	0.764	0.938	0.861	0.919	0.779	0.953	0.860	0.923	0.714	0.949	0.869	0.935	0.800	0.950	0.860	0.941
aive Bayes classifier																
aseline sum scores																
accuracy	0.633				0.670				0.672				0.634			
95% CI	0.606 - 0.660				0.643 - 0.697				0.644 - 0.699				0.604 - 0.663			
p value [acc > NIR)	0.996				>0.999				1.000				1.000			
balanced accuracy	0.738	0.582	0.606	0.635	0.723	0.538	0.649	0.644	0.717	0.591	0.629	0.630	0.708	0.550	0.602	0.592
sensitivity	0.785	0.259	0.343	0.343	0.796	0.140	0.425	0.355	0.785	0.228	0.381	0.355	0.754	0.154	0.320	0.323
specificity	0.691	0.904	0.870	0.927	0.650	0.936	0.873	0.933	0.649	0.953	0.878	0.905	0.663	0.947	0.885	0.861
positive predictive value	0.837	0.152	0.341	0.356	0.851	0.099	0.369	0.342	0.867	0.206	0.331	0.213	0.865	0.133	0.320	0.135
negative predictive value	0.616	0.948	0.871	0.923	0.560	0.956	0.896	0.937	0.508	0.959	0.899	0.951	0.484	0.955	0.885	0.950
gistic regression																
aseline and sum scores																
accuracy	0.685				0.727				0.740				0.735			
95% CI	0.658 - 0.710				0.701 - 0.752				0.714 - 0.766				0.707 - 0.762			
P value [acc > NIR)	0.102				0.177				0.648				0.704			
balanced accuracy	0.671	0.515	0.603	0.589	0.613	0.501	0.575	0.567	0.575	0.516	0.549	0.549	0.584	0.504	0.539	0.533
sensitivity	0.912	0.037	0.305	0.219	0.958	0.018	0.190	0.150	0.956	0.035	0.135	0.118	0.961	0.019	0.107	0.092
specificity	0.429	0.993	0.901	0.960	0.268	0.984	0.960	0.984	0.194	0.997	0.962	0.980	0.206	0.989	0.972	0.973
positive predictive value	0.763	0.250	0.378	0.390	0.766	0.053	0.453	0.471	0.776	0.400	0.362	0.300	0.777	0.083	0.390	0.188
negative predictive value	0.709	0.939	0.869	0.912	0.719	0.953	0.871	0.922	0.602	0.951	0.875	0.939	0.647	0.950	0.865	0.941

ote. The *p* value denotes the one sided ANOVA statistic of accuracy (acc) compared with the No-Information Rate (NIR). When accuracy is smaller than NIR, the *p* value is 1.000. NIR was 0.668, 0.714, 744, and 0.742 for year 2-, 4-, 6-, and 9-year outcome, respectively. I denotes healthy. II denotes mood disorder. III denotes anxiety disorder. IV denotes comorbid mood/anxiety disorder.

pplementary Table 5 Predicting mental health status (categorical outcome) at 2-, 4-, 6-, and 9-year follow up using baseline and 1-year sum and item-scores as the independent variables (i.e., edictor Set D).

AUTO-SKLEARN	Outcome at Year 2				Outcome at Year 4				Outcome at Year 6				Outcome at Year 9			
ealth status	**I**	**II**	**III**	**IV**	**I**	**II**	**III**	**IV**	**I**	**II**	**III**	**IV**	**I**	**II**	**III**	**IV**
aseline sum scores and individual items																
accuracy	0.704				0.718				0.747				0.748			
95% CI	0.679 - 0.729				0.691 - 0.743				0.721 - 0.772				0.720 - 0.774			
p value [acc > NIR)	0.003				0.413				0.434				0.350			
balanced accuracy	0.660	0.500	0.595	0.624	0.540	0.500	0.524	0.538	0.561	0.500	0.523	0.534	0.539	0.500	0.519	0.534
sensitivity	0.948	0.000	0.239	0.299	0.980	0.000	0.067	0.084	0.983	0.000	0.071	0.079	0.992	0.000	0.047	0.077
specificity	0.371	1.000	0.950	0.949	0.099	0.999	0.980	0.992	0.139	1.000	0.974	0.989	0.086	1.000	0.991	0.991
positive predictive value	0.752	0.000	0.486	0.410	0.731	0.000	0.375	0.500	0.769	0.000	0.306	0.333	0.757	0.000	0.467	0.357
negative predictive value	0.780	0.938	0.864	0.920	0.667	0.953	0.857	0.917	0.741	0.949	0.868	0.937	0.793	0.950	0.860	0.941
aive Bayes classifier																
aseline sum scores and individual items																
accuracy	0.621				0.648				0.655				0.608			
95% CI	0.594 - 0.647				0.620 - 0.675				0.627 - 0.683				0.577 - 0.638			
p value [acc > NIR)	<0.999				1.000				1.000				1.000			
balanced accuracy	0.748	0.598	0.584	0.664	0.731	0.537	0.610	0.684	0.727	0.608	0.603	0.656	0.686	0.621	0.576	0.639
sensitivity	0.764	0.296	0.277	0.445	0.770	0.158	0.318	0.477	0.767	0.298	0.290	0.434	0.712	0.327	0.247	0.431
specificity	0.733	0.900	0.890	0.883	0.691	0.915	0.901	0.891	0.688	0.918	0.915	0.878	0.659	0.916	0.906	0.846
positive predictive value	0.852	0.164	0.331	0.310	0.862	0.085	0.361	0.300	0.877	0.162	0.352	0.205	0.857	0.170	0.308	0.158
negative predictive value	0.607	0.951	0.863	0.931	0.546	0.956	0.883	0.946	0.503	0.961	0.890	0.956	0.443	0.963	0.877	0.957
gistic regression																
aseline sum scores and individual items																
accuracy	0.657				0.694				0.708				0.696			
95% CI	0.631 - 0.683				0.668 - 0.720				0.681 - 0.735				0.667 - 0.724			
p value [acc > NIR)	0.804				0.941				0.997				1.000			
balanced accuracy	0.686	0.534	0.587	0.603	0.624	0.493	0.575	0.559	0.598	0.524	0.549	0.560	0.591	0.512	0.546	0.525
sensitivity	0.862	0.099	0.286	0.270	0.907	0.018	0.218	0.150	0.905	0.088	0.142	0.158	0.897	0.058	0.147	0.092
specificity	0.510	0.970	0.887	0.936	0.341	0.968	0.932	0.969	0.292	0.960	0.956	0.963	0.285	0.965	0.946	0.957
positive predictive value	0.780	0.178	0.332	0.333	0.775	0.026	0.358	0.320	0.788	0.104	0.338	0.235	0.783	0.081	0.314	0.125
negative predictive value	0.647	0.942	0.864	0.916	0.594	0.952	0.872	0.921	0.512	0.952	0.875	0.941	0.490	0.951	0.867	0.940

ote. The *p* value denotes the one sided ANOVA statistic of accuracy (acc) compared with the No-Information Rate (NIR). When accuracy is smaller than NIR, the *p* value is 1.000. NIR was 0.668, 714, 0.744, and 0.742 for year 2-, 4-, 6-, and 9-year outcome, respectively. I denotes healthy. II denotes mood disorder. III denotes anxiety disorder. IV denotes comorbid mood/anxiety disorder.


TRUE VALUES
Pedictor Set A
Predictor Set B
Predictor Set C
Predictor Set D
Healthy
Disorder
Logistic regression
Naïve Bayes
AUTO-SKLEARN
767 176 102 255
723 185 144 246
754 191 113 240
724 191 143 240
689 136 178 295
707 139 160 292
690 141 177 290
686 123 181 302
725 158 142 273
733 153 134 278
761 201 106 230
773 201 94 230
Healthy
Mood disorder
Anxiety disorder
Comorbidity
Logistic regression
Naïve Bayes
AUTO-SKLEARN
799 55 112 57 4 6 7 3 49 10 71 40 15 10 23 37
695 44 83 47 54 11 15 21 80 14 73 34 38 12 42 35
791 63 117 66 2 3 4 3 60 9 65 38 14 6 27 30
747 52 107 52 14 8 11 12 78 9 61 36 28 12 34 37
678 37 66 24 75 26 30 28 92 10 73 27 22 8 44 58
677 31 64 24 60 21 18 26 87 11 67 17 43 15 64 70
681 41 65 27 58 21 28 31 101 8 73 32 27 11 47 47
662 31 64 20 71 24 19 32 85 10 59 24 49 16 71 61
835 68 136 70 0 0 0 0 20 7 48 18 12 6 29 49
834 72 136 72 0 0 0 0 22 4 55 30 11 5 22 35
821 71 131 80 0 0 0 0 23 3 45 16 23 7 37 41
822 67 131 73 0 0 0 0 24 7 51 23 21 7 31 41

Supplementary Figure 1. Confusion Matrixes.

Upper confusion matrices depict the binary predictions, that is, (mentally) healthy or mood/anxiety disorder. The lower confusion matrices depict the categorical predictions, that is, (mentally) healthy, mood disorder, anxiety disorder, comorbid mood and anxiety disorder. The number in each cell describes the number of predicted diagnostic categories (*y*-axis) in relation to the true diagnostic categories (*x*-axis). The black borders depicts the correctly classified participants (i.e., true positive and true negative values). All predictor sets included baseline psychiatric diagnoses and demographic variables. Predictor Set A further includes baseline and 1-year follow-up sum scores. Predictor Set B additionally includes baseline and 1-year follow-up individual items. Predictor Set C includes baseline sum scores. Predictor Set D additionally includes individual items.