Psychiatry Research 299 (2021) 113823

Contents lists available at ScienceDirect

Psychiatry Research

journal homepage: www.elsevier.com/locate/psychres

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

Wessel A. van Eeden[a,*], Chuan Luo[b], Albert M. van Hemert[a], Ingrid V.E. Carlier[a], Brenda W. Penninx[c], Klaas J. Wardenaar[d], Holger Hoos[b], Erik J. Giltay[a]

[a] Department of Psychiatry, Leiden University Medical Center, Leiden, the Netherlands
[b] Leiden Institute of Advanced Computer Sciences, Leiden University, Leiden, the Netherlands
[c] Department of Psychiatry, Amsterdam Public Health Research Institute and Amsterdam Neuroscience, VU University Medical Center, and GGZ inGeest, Amsterdam, the Netherlands
[d] Department of Psychiatry, The University Medical Center Groningen, Groningen, the Netherlands

ARTICLE INFO

Keywords:
Psychiatry
Depression
Anxiety disorder
Machine learning
Logistic models
Epidemiologic methods
Regression analysis

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.

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 (McGorry, 2010). In clinical settings, most decision making is based on clinical-care guidelines and experience (Ægisdóttir et al., 2006). However, even experienced clinicians may ignore relevant information or may put too much emphasis on clinically salient cues (Odeh et al., 2006). Information on demographic characteristics and clinician-rated and self-reported measures are increasingly collected as part of routine outcome monitoring (ROM; Carlier et al., 2012), 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 (Ægisdóttir et al., 2006; Grove et al., 2000), particularly through modern machine learning (ML) approaches (Johnson et al., 2016).

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 (Iniesta et al., 2016; Wang et al., 2018). Most clinical data thus far have been analyzed by selecting only specific putative predictors. It is possible that more complex (including nonlinear and higher

* Corresponding author at: Department of Psychiatry, PO Box 9600, 2300 RC, Leiden, the Netherlands
E-mail address: W.A.van_Eeden@lumc.nl (W.A. van Eeden).

https://doi.org/10.1016/j.psychres.2021.113823
Received 4 October 2020; Accepted 20 February 2021
Available online 24 February 2021

dimensional) patterns exist in the data, which can efficiently be detected when analyzing all available data simultaneously using ML (Chekroud et al., 2016; Hahn et al., 2016). These approaches are able to examine huge numbers of potential predictors in an unbiased manner while preventing overfitting (Hastie et al., 2009).

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% confidence interval [CI] .77–.87; Lee et al., 2018). 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; Dinga et al., 2018). 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% ($p < .05$; Bokma et al., 2020). Clinical severity measures were the most important predictor variables, which is in line with previous reports (Bokma et al., 2020; Dinga et al., 2018; Lee et al., 2018). Although these studies seem promising, recently published papers have demonstrated only limited added value of ML over traditional regression analyses (Christodoulou et al., 2019; van Mens et al., 2020). 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 (Belsher et al., 2019; Kessler et al., 2017). Despite the increasing number of publications in this field, ML has yet to move towards clinical application (Tran et al., 2019).

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 (Waring et al., 2020). 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 (Feurer et al., 2015; Feurer et al., 2019).

We aimed to study and to compare the performance of traditional multinominal logistic regression, a basic probabilistic ML algorithm (naïve Bayesian classifier; Jayant and Safari, 2020) 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.

2. Methods

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 (Penninx et al., 2008). 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%; Penninx et al., 2008) 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 (Penninx et al., 2008). 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.

2.2. Measures

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; Marks and Mathews, 1979), the Beck's Anxiety Inventory (BAI; Beck et al., 1988), and the Inventory of Depressive Symptomatology (IDS-SR; Rush et al., 1996). 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.

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 (Penninx et al., 2008; Wittchen, 1994) and may be considered a gold standard for psychiatric diagnostic classification (Haro et al., 2006; Kessler et al., 2009).

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.

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 (Menard, 2002), naïve Bayes classifier (Jayant and Safari, 2020), and Auto-sklearn (Feurer et al., 2015). 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 Feurer et al., 2019). Previous research shows that the classification performance is often much better than using standard selection/hyperparameter optimization methods (Feurer et al., 2015), and researchers believe Auto-sklearn to be a promising system for use by both ML novices and experts (Feurer et al., 2019). 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 (Feurer et al., 2019). More details about Auto-sklearn can be found elsewhere (Feurer et al., 2015; Feurer et al., 2019; 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 (Jayant and Safari, 2020). 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 (Menard, 2002). We used the `R` package `nnet` (R Foundation for Statistical Computing, Vienna, Austria, 2016. https://www.R-project.org/; Ripley et al., 2016).

We computed all models by randomly splitting (50:50) the dataset into a training and a test dataset using `Scikit-learn` data split (Pedregosa et al., 2011). 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 analyses. 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 (Rossum, 1995). For logistic regression, `R` was used (R Foundation for Statistical Computing, Vienna, Austria, 2016. https://www.R-project.org/; Ripley et al., 2016).

3. Results

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 ($n = 742$).

3.2. Prediction of health status as binary outcome

Figs. 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. Fig. 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 Fig. 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 Fig. 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

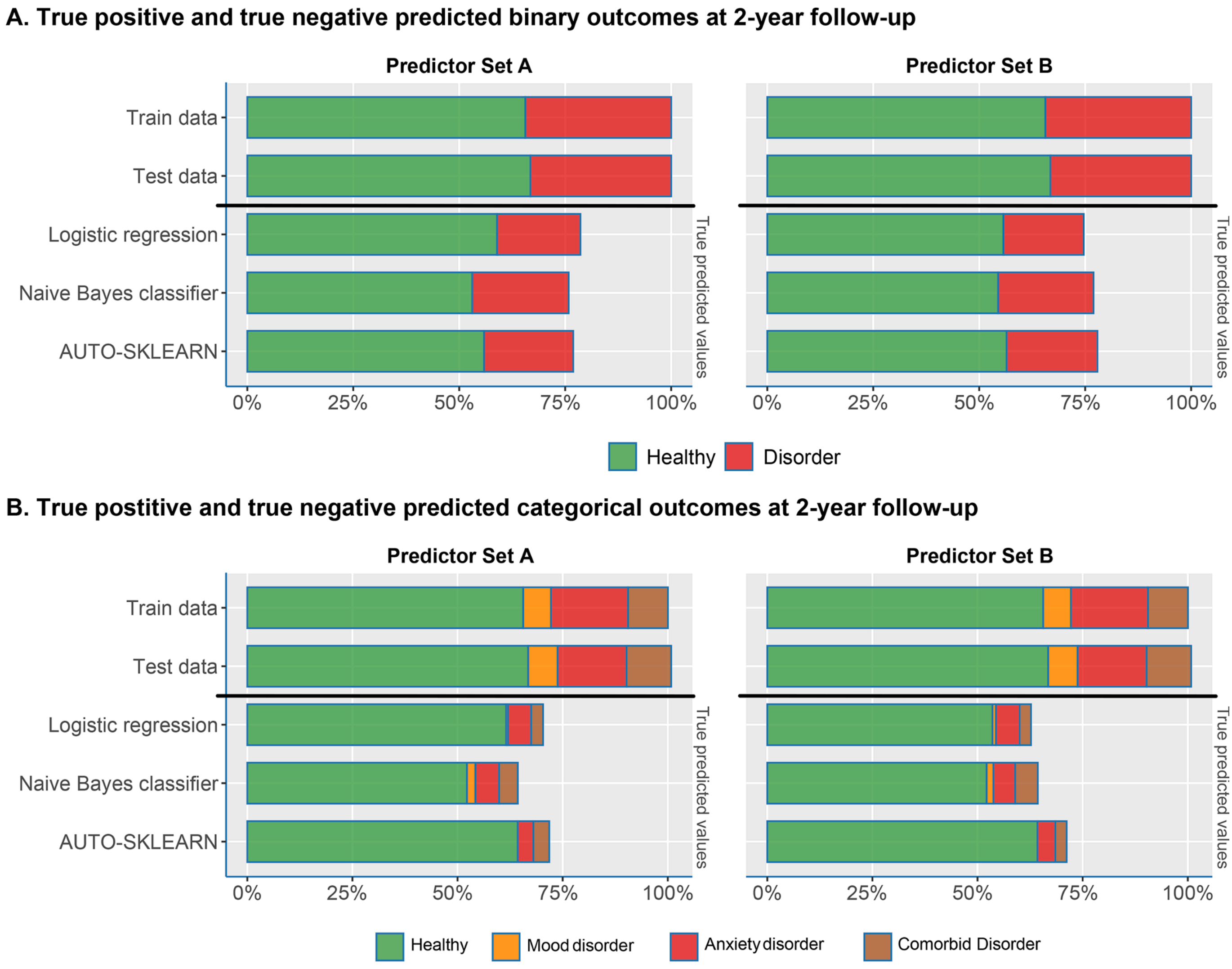

Fig. 1. Percentages of train and test dataset values, as well as those correctly predicted at 2-year follow up, using the three data models. All predictor sets included baseline psychiatric diagnoses and demographic variables. Predictor Set A further includes baseline and 1-year follow-up sum scores.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

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 Fig. 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.

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 Figs. 1, 3, and 4 and in the supplementary material Fig. 1 and Tables 4 and 5. Fig. 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 Fig. 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 Fig. 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 Figs. 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 Fig. 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]; $p < .001$;

	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*

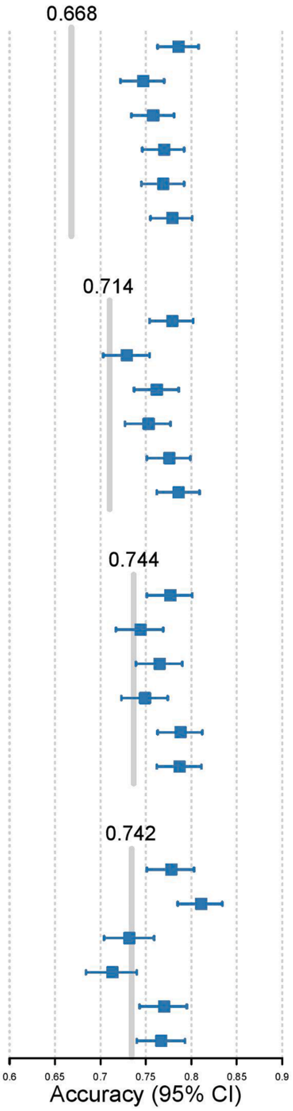

Fig. 2. Predicting health status (binary outcome) at 2-, 4-, 6-, and 9-year follow up. All predictor sets included baseline psychiatric diagnoses and demographic variables. Predictor Set A further includes baseline and 1-year follow-up sum scores. Predictor Set B additionally includes baseline and 1-year 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

see Fig. 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.

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

	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	

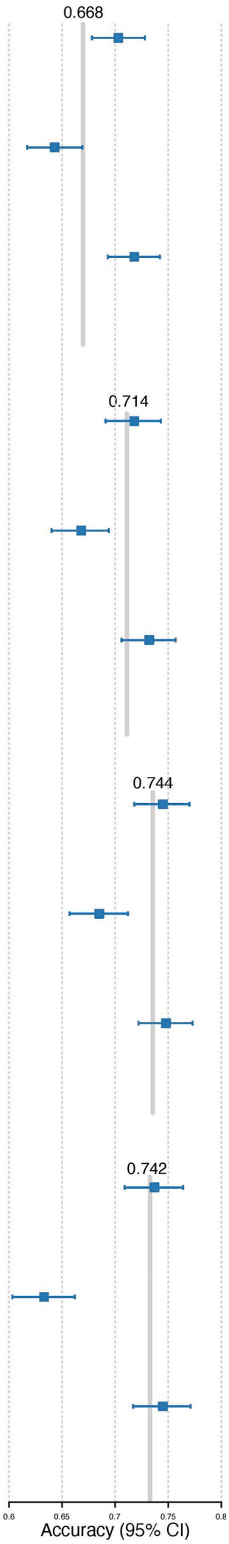

Fig. 3. Predicting health status (multinominal outcome) at 2-, 4-, 6-, and 9-year follow up with baseline and 1-year sum scores (predictor Set A). All predictor sets included baseline psychiatric diagnoses and demographic variables. Predictor Set A further includes baseline and 1-year follow-up sum scores. 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	

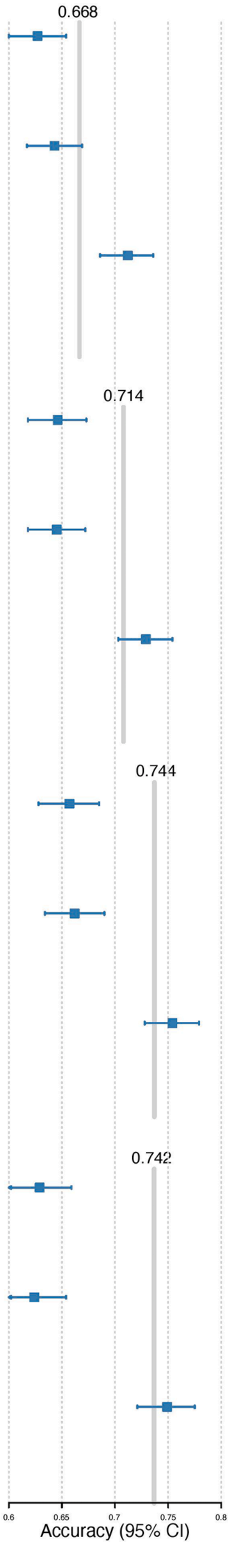

Fig. 4. Predicting health status (multinominal outcome) at 2-, 4-, 6-, and 9-year follow up with baseline and 1-year sum scores and individual item-scores (predictor Set B). All predictor sets included baseline psychiatric diagnoses and demographic variables. Predictor Set B further includes baseline 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

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 (Bokma et al., 2020; Chekroud et al., 2016; Dinga et al., 2018; Kessler et al., 2016; Lee et al., 2018; Nie et al., 2018). 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% (Dinga et al., 2018). 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% (Bokma et al., 2020). 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 (Bokma et al., 2020; Dinga et al., 2018). 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 (Chekroud et al., 2016; Kessler et al., 2016). 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 (Nie et al., 2018). 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 (Chekroud et al., 2016; Kessler et al., 2016). Although level of accuracy was higher for ML, this difference was not found to be significant in either study (Chekroud et al., 2016; Kessler et al., 2016). 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 (Tran et al., 2019). 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., Christodoulou et al., 2019; Desai et al., 2020; Frizzell et al., 2017). However, it is possible that ML does outperform traditional methods when more complex (large) datasets are used (Iniesta et al., 2016; Wang et al., 2018). 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 (Lee et al., 2018). 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 (Ægisdóttir et al., 2006). 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 (Ægisdóttir et al., 2006). With a statistical approach, statistical methods are applied on objectively measured variables in order to make predictions and prognoses based on probabilities (Ægisdóttir et al., 2006). Two meta-analyses demonstrated that statistical approaches were more accurate than clinical methods (Ægisdóttir et al., 2006; Grove et al., 2000). 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 (Waring et al., 2020). Moreover, Auto-sklearn demonstrated that it can perform even under rigid time and computational resource constraints (Feurer et al., 2015). Automated ML is already demonstrating its usefulness in healthcare practice (Waring et al., 2020).

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 (Regier et al., 2013). Despite our relatively large sample size, our analyses 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 (Chekroud et al., 2016; Nie et al., 2018). Although we made use of a training and testing dataset, it is possible that the results from the ML methods and regression analyses 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 (Regier et al., 1998). 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 (Jacobson and Newman, 2017; Shorter and Tyrer, 2003). Sixth, ML may have more added value when the dataset is more complex, such as imaging or genetic data (Iniesta et al., 2016; Lee et al., 2018; Wang et al., 2018). 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.

Author statement: contributors

BP is principal investigator of NESDA. Author WE wrote and performed the methodology, and wrote and edited the manuscript. CL performed the methodology. WE had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. EG contributed by frequent supervision of the writing of this manuscript. EG and HH contributed by supervision of the methodology. CL, EG, AH, IC, BP, KW, and HH contributed by review and editing of the early stages as well as final stages of the manuscript. All authors had access to the data. All authors have approved the final manuscript.

Role of the funding source

The funding source had no role in the design of this study, it's execution, analyses, interpretation of the data, or decision to submit results.

Declaration of Competing Interests

BP has received (non-related) research funding from Boehringer Ingelheim and Jansen Research. All remaining authors report no financial interests or potential conflicts of interest.

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) .

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.psychres.2021.113823.

References

Ægisdóttir, S., White, M.J., Spengler, P.M., Maugherman, A.S., Anderson, L.A., Cook, R. S., Nichols, C.N., Lampropoulos, G.K., Walker, B.S., Cohen, G., 2006. The meta-analysis of clinical judgment project: fifty-six years of accumulated research on clinical versus statistical prediction. Counsel. Psychol. 34 (3), 341–382.

Beck, A.T., Epstein, N., Brown, G., Steer, R.A., 1988. An inventory for measuring clinical anxiety: psychometric properties. J. Consult. Clin. Psychol. 56 (6), 893.

Belsher, B.E., Smolenski, D.J., Pruitt, L.D., Bush, N.E., Beech, E.H., Workman, D.E., Morgan, R.L., Evatt, D.P., Tucker, J., Skopp, N.A., 2019. Prediction models for suicide attempts and deaths: a systematic review and simulation. JAMA Psychiatry 76 (6), 642–651.

Bokma, W.A., Zhutovsky, P., Giltay, E.J., Schoevers, R.A., Penninx, B.W., van Balkom, A. L., Batelaan, N.M., van Wingen, G.A., 2020. Predicting the naturalistic course in anxiety disorders using clinical and biological markers: a machine learning approach. Psychol. Med. 1–11.

Carlier, I.V., Meuldijk, D., Van Vliet, I.M., Van Fenema, E., Van der Wee, N.J., Zitman, F. G., 2012. Routine outcome monitoring and feedback on physical or mental health status: evidence and theory. J. Eval. Clin. Pract. 18 (1), 104–110.

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

Christodoulou, E., Jie, M., Collins, G.S., Steyerberg, E.W., Verbakel, J.Y., van Calster, B., 2019. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J. Clin. Epidemiol.

Desai, R.J., Wang, S.V., Vaduganathan, M., Evers, T., Schneeweiss, S., 2020. Comparison of machine learning methods with traditional models for use of administrative claims with electronic medical records to predict heart failure outcomes. JAMA Netw. Open 3 (1) e1918962-e1918962.

Dinga, R., Marquand, A.F., Veltman, D.J., Beekman, A.T., Schoevers, R.A., van Hemert, A.M., Penninx, B.W., Schmaal, L., 2018. Predicting the naturalistic course of depression from a wide range of clinical, psychological, and biological data: a machine learning approach. Transl. Psychiatry 8 (1), 241.

Feurer, M., Klein, A., Eggensperger, K., Springenberg, J., Blum, M., Hutter, F., 2015. Efficient and robust automated machine learning. Adv. Neural Inf. Process. Syst. 2962–2970.

Feurer, M., Klein, A., Eggensperger, K., Springenberg, J.T., Blum, M., Hutter, F., 2019. Auto-sklearn: efficient and robust automated machine learning. Automated Machine Learning. Springer, pp. 113–134.

Frizzell, J.D., Liang, L., Schulte, P.J., Yancy, C.W., Heidenreich, P.A., Hernandez, A.F., Bhatt, D.L., Fonarow, G.C., Laskey, W.K., 2017. Prediction of 30-day all-cause readmissions in patients hospitalized for heart failure: comparison of machine learning and other statistical approaches. JAMA Cardiol. 2 (2), 204–209.

Grove, W.M., Zald, D.H., Lebow, B.S., Snitz, B.E., Nelson, C., 2000. Clinical versus mechanical prediction: a meta-analysis. Psychol. Assess. 12 (1), 19.

Hahn, T., Nierenberg, A., Whitfield-Gabrieli, S., 2016. Predictive analytics in mental health: applications, guidelines, challenges and perspectives. Mol. Psychiatry 22 (1), 37.

Haro, J.M., Arbabzadeh-Bouchez, S., Brugha, T.S., De Girolamo, G., Guyer, M.E., Jin, R., Lepine, J.P., Mazzi, F., Reneses, B., Vilagut, G., 2006. Concordance of the composite international diagnostic interview version 3.0 (CIDI 3.0) with standardized clinical assessments in the WHO world mental health surveys. Int. J. Methods Psychiatr. Res. 15 (4), 167–180.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Series in Statistics. Springer New York.

Iniesta, R., Stahl, D., McGuffin, P., 2016. Machine learning, statistical learning and the future of biological research in psychiatry. Psychol. Med. 46 (12), 2455–2465.

Jacobson, N.C., Newman, M.G., 2017. Anxiety and depression as bidirectional risk factors for one another: a meta-analysis of longitudinal studies. Psychol. Bull. 143 (11), 1155.

Jayant, A., Safari, a.O.R.M.C., 2020. Data Science and Machine Learning Series: Naive Bayes Classifier Advanced Concepts. Technics Publications.

Johnson, A.E., Ghassemi, M.M., Nemati, S., Niehaus, K.E., Clifton, D.A., Clifford, G.D., 2016. Machine learning and decision support in critical care. Proceed. IEEE. Inst. Electr. Electron. Eng. 104 (2), 444.

Kessler, R.C., Avenevoli, S., Green, J., Gruber, M.J., Guyer, M., He, Y., Jin, R., Kaufman, J., Sampson, N.A., Zaslavsky, A.M., 2009. National comorbidity survey replication adolescent supplement (NCS-A): III. Concordance of DSM-IV/CIDI diagnoses with clinical reassessments. J. Am. Acad. Child Adolesc. Psychiatry 48 (4), 386–399.

Kessler, R.C., Hwang, I., Hoffmire, C.A., McCarthy, J.F., Petukhova, M.V., Rosellini, A.J., Sampson, N.A., Schneider, A.L., Bradley, P.A., Katz, I.R., 2017. Developing a practical suicide risk prediction model for targeting high-risk patients in the veterans health administration. Int. J. Methods Psychiatr. Res. 26 (3), e1575.

Kessler, R.C., van Loo, H.M., Wardenaar, K.J., Bossarte, R.M., Brenner, L.A., Cai, T., Ebert, D.D., Hwang, I., Li, J., de Jonge, P., 2016. Testing a machine-learning algorithm to predict the persistence and severity of major depressive disorder from baseline self-reports. Mol. Psychiatry 21 (10), 1366.

Lee, Y., Ragguett, R.-M., Mansur, R.B., Boutilier, J.J., Rosenblat, J.D., Trevizol, A., Brietzke, E., Lin, K., Pan, Z., Subramaniapillai, M., 2018. Applications of machine learning algorithms to predict therapeutic outcomes in depression: a meta-analysis and systematic review. J. Affect. Disord.

Marks, I.M., Mathews, A.M., 1979. Brief standard self-rating for phobic patients. Behav. Res. Ther. 17 (3), 263–267.

McGorry, P.D., 2010. Risk syndromes, clinical staging and DSM V: new diagnostic infrastructure for early intervention in psychiatry. Schizophrenia Res. 120 (1), 49–53.

Menard, S., 2002. Applied Logistic Regression Analysis. Sage.

Nie, Z., Vairavan, S., Narayan, V.A., Ye, J., Li, Q.S., 2018. Predictive modeling of treatment resistant depression using data from STAR*D and an independent clinical study. PLoS One 13 (6), e0197268.

Odeh, M.S., Zeiss, R.A., Huss, M.T., 2006. Cues they use: clinicians' endorsement of risk cues in predictions of dangerousness. Behav. Sci. Law 24 (2), 147–156.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., 2011. Scikit-learn: machine learning in Python. the. J. Mach. Learn. Res. 12, 2825–2830.

Penninx, B.W., Beekman, A.T., Smit, J.H., Zitman, F.G., Nolen, W.A., Spinhoven, P., Cuijpers, P., De Jong, P.J., Van Marwijk, H.W., Assendelft, W.J., Van Der Meer, K., Verhaak, P., Wensing, M., De Graaf, R., Hoogendijk, W.J., Ormel, J., Van Dyck, R., Consortium, N.R., 2008. The Netherlands study of depression and anxiety (NESDA): rationale, objectives and methods. Int. J. Methods Psychiatr. Res. 17 (3), 121–140.

Regier, D.A., Kaelber, C.T., Rae, D.S., Farmer, M.E., Knauper, B., Kessler, R.C., Norquist, G.S., 1998. Limitations of diagnostic criteria and assessment instruments for mental disorders: implications for research and policy. Arch. Gener. Psychiatry 55 (2), 109–115.

Regier, D.A., Kuhl, E.A., Kupfer, D.J., 2013. The DSM-5: classification and criteria changes. World Psychiatry 12 (2), 92–98.

Ripley, B., Venables, W., Ripley, M.B., 2016. Package 'nnet'. R Package Version, 7-3.

Rossum, G.V., 1995. Python tutorial, technical report CS-R9526. Centrum voor Wiskunde en Informatica (CWI). The Netherlands, Amsterdam.

Rush, A.J., Gullion, C.M., Basco, M.R., Jarrett, R.B., Trivedi, M.H., 1996. The inventory of depressive symptomatology (IDS): psychometric properties. Psychol. Med. 26 (3), 477–486.

Shorter, E., Tyrer, P., 2003. Separation of anxiety and depressive disorders: blind alley in psychopharmacology and classification of disease. BMJ 327 (7407), 158–160.

Tran, B.X., McIntyre, R.S., Latkin, C.A., Phan, H.T., Vu, G.T., Nguyen, H.L.T., Gwee, K.K., Ho, C.S., Ho, R., 2019. The current research landscape on the artificial intelligence application in the management of depressive disorders: a bibliometric analysis. Int. J. Environ. Res. Public Health 16 (12), 2150.

van Mens, K., de Schepper, C., Wijnen, B., Koldijk, S.J., Schnack, H., de Looff, P., Lokkerbol, J., Wetherall, K., Cleare, S., O'Connor, R.C., 2020. Predicting future suicidal behaviour in young adults, with different machine learning techniques: a population-based longitudinal study. J. Affect. Disord.

Wang, Y., Kung, L., Byrd, T.A., 2018. Big data analytics: understanding its capabilities and potential benefits for healthcare organizations. Technol. Forecast. Soc. Change 126, 3–13.

Waring, J., Lindvall, C., Umeton, R., 2020. Automated machine learning: review of the state-of-the-art and opportunities for healthcare. Artif. Intel. Med., 101822

Wittchen, H.-U., 1994. Reliability and validity studies of the WHO-composite international diagnostic interview (CIDI): a critical review. J. Psychiatr. Res. 28 (1), 57–84.