SUPPLEMENTARY MATERIAL

Validation of Random Forest Machine Learning Models to Predict Dementia-Related Neuropsychiatric Symptoms in Real-World Data

Methods and Results

Criteria for the diagnosis of dementia in the population electronic registry

The search for diagnoses was carried out following validated criteria with a positive predictive value of 95.1% and a negative value of 99.4%. The dementia codes included were in ICD9 (290xx; 2941x; 331xx) and ICD10 (F01.5x; F02.8x; F03.9x; G30.0x; G30.1x; G30.8x; G30.9x; G31.0x; G31.1x; G31.2x; G31.8x; G31.9x). The dementia codes with NPS used were searched in CIE9 (294.11, 290.11, 290.12, 290.13, 290.2x, 290.3, 290.41, 290.42, 290.43, 290.9) and CIE10 (F02.81, F03.90, F05). ICD-10 was added to the searches because from 1 January 2016, it is the official ICD adopted by the national health system. The identification of dementia also included the prescription of specific drugs for Alzheimer's disease (ATC Group N06D) such as acetylcholinesterase inhibitors (donepezil, rivastigmine, and galantamine), memantine, and rivastigmine.

Resources used in the review of electronic health records (EHRs)

To review the EHRs, the technician employed 1100 hours which means a mean time for each case of 16.5 minutes.

Correlation between presence of neuropsychiatric symptoms (NPS) in the EHR and Neuropsychiatric Inventory (NPI) score

We explored the correlation between presence of NPS in the EHR and NPI score in a small sample of patients (50 cases) diagnosed with dementia and living in the community. For this purpose, the Spanish version of the NPI [1] was administered and the EHR was searched for the presence of NPS.

We found that 25 had a record of the presence in their EHR of psychotic symptoms and 19 of depressive symptoms. Analyzing the NPI scores, we found that near all positive patients (25 with psychotic symptoms and 18 with depressive symptoms) scored four or more in the NPI domains correlating with the presence of depressive cluster symptoms (Anxiety, Depression/Dysphoria, Apathy/Indifference) and psychotic cluster symptoms (Delusions, Hallucinations, Agitation/Aggression, Irritability/Lability). In addition, the NPI identified seven further patients with psychiatric symptoms and ten more with depressive symptoms.

This small sample analysis does not mean a thorough validation study but does indicate that it is possible to identify the need for an intervention to manage the symptoms by reviewing EHRs with less sensitivity than using the NPI score but not less specificity.

Supplementary Table 1. Features of patients in the small sample.

	Number	
Sex (women)	36	72%
Age ≥85 y	30	60%
EHR Depressive cluster	19	38%
EHR Psychotic cluster ≥ 4	25	50%
NPI Depressive cluster ≥ 4	28	56%
NPI Psychotic cluster	33	66%
Total	50	100%

Supplementary Table 2. Correlation between presence of neuropsychiatric symptoms in the electronic health record and Neuropsychiatric Inventory (NPI).

		NPI score		
Depressive cluster		Positive	Negative	Total
EHR	Positive	18	1	19
	Negative	10	21	31
	Total	28	22	50
Psychotic cluster		Positive	Negative	Total
EHR	Positive	26	0	26
	Negative	7	17	24
	Total	33	17	50

Supplementary Table 3. Diagnostic features of the presence of symptoms in the electronic health record.

		95% Confidence Intervals	
Depressive cluster	Value	Lower	Upper
Prevalence	56.00%	41.35%	69.73%
Patients correctly diagnosed	78.00%	63.67%	88.01%
Sensitivity	64.29%	44.11%	80.69%
Specificity	95.45%	75.12%	99.76%
Positive predictive value	94.74%	71.89%	99.72%
Negative predictive value	67.74%	48.54%	82.68%
		95% Confidence Intervals	
Psychotic cluster	Value	Lower	Value
Prevalence	66.00%	51.14%	78.41%
Patients correctly diagnosed	86.00%	72.64%	93.72%
Sensitivity	78.79%	60.60%	90.37%
Specificity	100.00%	77.08%	99.46%
Positive predictive value	100.00%	83.98%	99.65%
Negative predictive value	70.83%	48.75%	86.56%

Data preprocessing

The dataset contained instances with 68 variables belonging to 4,003 patients, recorded as repeated-measures data. Each one of the data instances referred to a new prescription or change in a medication for a specific patient. Therefore, the dataset was preprocessed to transform it in such a way that it contained a single summarizing instance per person. In order to achieve this goal, the summarizing variables shown in Supplementary Table 1 were derived from the initial variables per patient, while some irrelevant features for the current study were rejected.

Supplementary Table 4. Description of variables included in the dataset.

Group		Variable
Response variable, NPS *Does the patient's EHR contain a record of symptoms sometime? (Yes/No)*		- Depressive-cluster NPS - Psychotic-cluster NPS
Basic information		- Anonymized Patient ID - Nursing home - Age - Sex
Medication	**Prescription history** *Has the patient taken any of the following medications at some point? (Yes/No)*	- Antidepressant treatment (any) - Antipsychotic treatment (any) - Specific medication (based on 31 ATCs: N05AA01, N05AA02, N05AD01, N05AH02, N05AH03, N05AH04, N05AL01, N05AL03, N05AN01, N05AX08, N05AX12, N05AX13, N06AA04, N06AA09, N06AA10, N06AA21, N06AB03, N06AB04, N06AB05, N06AB06, N06AB08, N06AB10, N06AX03, N06AX05, N06AX11, N06AX12, N06AX16, N06AX21, N06AX22, N06AX23, N06AX26)
	Details in the medication prescription history *How many prescriptions/ changes of the following types are registered for this patient?*	- No. of antidepressant treats.: Number of distinct antidepressant treatments prescribed - No. of changes from antipsychotic to antidepressant: Number of times patient has changed from an antipsychotic treatment to an antidepressant treatment - No. of antipsychotic treats.: Number of distinct antipsychotic treatments prescribed - No. of changes from antidepressant to antipsychotic: Number of times patient has changed from an antidepressant treatment to an antipsychotic treatment - Sedation level: maximum sedative level prescribed, categorized as none/minimal/mild/ moderate/deep
Comorbidities *Has the patient been diagnosed with any of the following at some point? (Yes/No)*		- Diabetes mellitus - Dyslipidemia - Parkinson's disease - Hypertension - Thyroid disease - Stroke - Cardiovascular disease - Traumatic brain injury

Supplementary Table 5. Sedative capacity of all the drugs included in the dataset.

	ATC	Drug name	Sedative capacity (0: none; 1: minimal; 2: mild; 3: moderate; 4: deep)
Antipsychotic drugs	N05AA01	Chlorpromazine	3
	N05AA02	Levomepromazine	3
	N05AD01	Haloperidol	2
	N05AH02	Clozapine	4
	N05AH03	Olanzapine	3
	N05AH04	Quetiapine	3
	N05AL01	Sulpiride	1
	N05AL03	Tiapride	1
	N05AN01	Lithium	
	N05AX08	Risperidone	2
	N05AX12	Aripiprazole	1
	N05AX13	Paliperidone	2
Antidepressant drugs	N06AA04	Clomipramine	2
	N06AA09	Amitriptyline	2
	N06AA10	Nortriptyline	2
	N06AA21	Maprotiline	2
	N06AB03	Fluoxetine	0
	N06AB04	Citalopram	0
	N06AB05	Paroxetine	0
	N06AB06	Sertraline	0
	N06AB08	Fluvoxamine	1
	N06AB10	Escitalopram	0
	N06AX03	Mianserin	2
	N06AX05	Trazodone	2
	N06AX11	Mirtazapine	2
	N06AX12	Bupropion	0
	N06AX16	Venlafaxine	0
	N06AX21	Duloxetine	0
	N06AX22	Agomelatine	
	N06AX23	Desvenlafaxine	0
	N06AX26	Vortioxetine	0

Machine learning

Random forests are a powerful machine learning (ML) technique that have shown great efficiency in many fields. Among their advantages, we could highlight that as a bagging technique, they typically outperform many other common ML algorithms, as they act through a majority voting of many "weak" learners. At the same time, this combination of multiple trees offers a low bias and low variance in the results. Furthermore, it can handle all types of features (binary, categorical, or numerical) without the need to apply cumbersome preprocessing techniques that could lead to errors, and what is more, deal with multicollinearity issues particularly well. It is also robust to outliers and non-linearities, and can handle unbalanced data. Recent literature in data science in the biomedical field demonstrates how efficient these algorithms can be [2, 3]. Their intrinsic calculation of "feature importance" makes them especially interesting for model interpretation and the search for biomarkers in the biomedical field, making them a more convenient option than the currently promising deep network techniques for applications where the goal is not only to make good predictions, but to learn from them. The amount of training data required for random forests compared to deep learning models also makes the former a good choice for the current work. The package used to fit the random forest models and the ROC curves was R version 3.6.1. and its libraries Caret, randomForest, and pROC.

Deep learning approaches were not used to avoid their "black box" functioning. Deep learning (DL) techniques for data analysis have gained popularity in recent years, not undeservedly, as they have shown great capabilities, especially in the analysis of big image databases. Nevertheless, when large amounts of training data are not available or when explainable models are required, both being relevant issues in this study, the choice of DL is no

longer so clear. Indeed, classic ML algorithms have shown to outperform deep learning algorithms in small- and medium-sized datasets. Ensemble-based alternatives such as random forest are being widely used for advances in the paradigm of explainable artificial intelligence and there is growing interest in the healthcare sector due to the particular need to understand the decisions made by the models and the desire to generate knowledge about specific diseases in the search for new biomarkers.

Implementation of random forests

We assessed various ML models before opting for the Random Forest algorithm. These included Naïve Bayes, Logistic Regression, and Support Vector Machines (both with lineal and RBF kernels), among others. The spot check was also repeated for several iterations (for different numbers of predictors). No model stood out in an exaggerated way from the others, but the Random Forests gave the best ROC and sensitivity values in most cases, with the add-on of the interpretability capabilities, due to the intrinsic calculation of feature importance, as well as to the lower complexity when compared with non-linear SVMs (RBF), as the linear version performed significantly worse.

The sample (N) was randomly divided into a training set ($N_1 = 0.75 * N$) and a test set ($N_2 = 0.25 * N$). It was checked that patient characteristics did not differ between training and test sets (Supplementary Table 4). The ML approach was applied to build predictive models. After assessing various ML algorithms, the random forest was selected for the current study [4]. The random forest algorithm [5] is a stochastic ensemble method that uses bagging, a combination of bootstrapping and aggregation of weak learners, more specifically, decision trees, seeking to detect patterns in data and use these to predict outcomes, in our case, NPS [6]. The procedure

relies on training a preset number of trees (500 by default). Then, only a subpopulation of the samples available, chosen by bootstrapping, is employed for training. By default, the size of the subpopulations is the same as that of the original population, and hence, the expectation is that 67% of the samples will be included in each subpopulation, the rest being simply replicates. At each node of each tree, only a subset of the original features is randomly selected for deciding how to split into two branches. By default, √M are used, where M is the number of features. In this case, a grid search for the parameter fitting was performed. The values we have tried included √M and M, as well as some values smaller than √M and between it and M. For the values below √M, the results were quite worse (>10% worse) than for √M. Some values over √M, gave results slightly better (but below 2% better). Lastly, from one value on, a deep degradation was visible, due to certain overfitting effect. Therefore, despite of not being the best possible value if we strictly focus on AUC, it was preferred because of somehow reducing model dependency on the train data, hoping that the future predictions on unseen data would be better. When the model is applied to a new sample for validation, the decision is obtained by combining all individual tree decisions by aggregation (majority voting in classification problems and averaging in regression problems). The rationale behind random forest is that the bias of the full ensemble of trees is equivalent to the bias of each single tree, whereas the variance is much smaller [5].

In the training set, we followed a stepwise process beginning with baseline models whose performance was improved by adding other explanatory variables in an iterative way to test their contribution. The iterations and modelling variables tested in each model are summarized in Table 2. Mean decrease accuracy was used to assess the relative feature importance of the variables in the models [7]. This technique computes the accuracy of the trees that build the

model for the out-of-the bag sample of each tree. Then, for each variable, it permutes the values of the variables one after another and measures how much the accuracy changes. Any decrease in accuracy resulting from this permutation is averaged over all trees, and used as a measure of the importance of each variable in the random forest model.

All the predictive models have been evaluated using a k-fold cross validation approach, with k =10 and 10 repetitions. This evaluation technique consists of splitting the number of data instances available into k subsets, such that the model is trained and tested k times, using k-1 subsets for the training each time, and the remaining subset for the testing, until all subsets have been used once for testing. As the random seed for splitting the dataset may introduce a small variation in the results, the repeated approach makes it possible to obtain more stable and realistic results by computing the average metrics for 10 repetitions using 10 different seeds to split the dataset. The main advantage of this evaluation technique is that it maximizes the availability of data for training the models, as it allows all the data instances to be used both for training and validation purposes in different iterations. In addition, it gives accurate estimates of the performance of the prediction models for unseen data. The same process was carried out separately for the psychotic and depressive symptom models for which discriminatory power was assessed.

Discriminatory power refers to the ability of a prediction model to distinguish between two outcome classes. In order to evaluate the classification ability of the models, the following statistics were calculated for each model: the area under the receiver operating curve (AUC), sensitivity, specificity, accuracy, no-information rate and Kappa index. The AUC gives an overview of a model's ability to discriminate between positive and negative classes, independently of their prevalence, and is therefore suitable for imbalanced datasets. Sensitivity

or the true positive rate (TPR) is defined as the number of cases from the positive class that were predicted correctly by the model, while specificity or the true negative rate refers to the number of cases from the negative class that were actually predicted as negative. The no-information rate is the accuracy that can be achieved without using any model. Accuracy means the proportion of correct classifications by the model. The Kappa index measures the agreement between two approaches to classify mutually exclusive categories, agreement being characterized as slight (for values of 0–0.20), fair (0.21–0.40), moderate (0.41–0.60), or substantial (0.61–0.80) [8].

Supplementary Table 6. Treatments of patients disaggregated by psychotic and depressive disorder according to health record review.

	Psychotic symptoms in the EHR		Depressive symptoms in the EHR		
	Yes N=2307	No N=1696	Yes N=2356	No N=1647	Total N=4003
Antipsychotic treatment	1628 (70.6%)	467 (27.5%)	1337 (56.7%)	758 (46.0%)	2095 (52.3%)
Antidepressant treatment	1731 (75.0%)	988 (58.3%)	1890 (80.2%)	829 (50.3%)	2719 (67.9%)
Trazodone	1114 (48.3%)	405 (23.9%)	1007 (42.7%)	512 (31.1%)	1519 (37.9%)
Paroxetine	182 (7.89%)	138 (8.14%)	251 (10.7%)	69 (4.19%)	320 (7.99%)
Citalopram	160 (6.94%)	114 (6.72%)	220 (9.34%)	54 (3.28%)	274 (6.84%)
Mirtazapine	299 (13.0%)	195 (11.5%)	390 (16.6%)	104 (6.31%)	494 (12.3%)
Duloxetine	207 (8.97%)	123 (7.25%)	284 (12.1%)	46 (2.79%)	330 (8.24%)
Amitriptyline	104 (4.51%)	76 (4.48%)	124 (5.26%)	56 (3.40%)	180 (4.50%)
Desvenlafaxine	56 (2.43%)	25 (1.47%)	71 (3.01%)	10 (0.61%)	81 (2.02%)
Fluoxetine	55 (2.38%)	53 (3.12%)	88 (3.74%)	20 (1.21%)	108 (2.70%)
Sertraline	254 (11.0%)	148 (8.73%)	329 (14.0%)	73 (4.43%)	402 (10.0%)
Escitalopram	385 (16.7%)	266 (15.7%)	545 (23.1%)	106 (6.44%)	651 (16.3%)
Mianserin	32 (1.39%)	26 (1.53%)	41 (1.74%)	17 (1.03%)	58 (1.45%)
Venlafaxine	135 (5.85%)	69 (4.07%)	169 (7.17%)	35 (2.13%)	204 (5.10%)
Bupropion	22 (0.95%)	15 (0.88%)	32 (1.36%)	5 (0.30%)	37 (0.92%)
Vortioxetine	31 (1.34%)	27 (1.59%)	54 (2.29%)	4 (0.24%)	58 (1.45%)
Agomelatine	23 (1.00%)	15 (0.88%)	34 (1.44%)	4 (0.24%)	38 (0.95%)
Nortriptyline	1 (0.04%)	1 (0.06%)	2 (0.08%)	0 (0.00%)	2 (0.05%)
Maprotiline	2 (0.09%)	2 (0.12%)	4 (0.17%)	0 (0.00%)	4 (0.10%)
Clomipramine	10 (0.43%)	3 (0.18%)	10 (0.42%)	3 (0.18%)	13 (0.32%)
Fluvoxamine	5 (0.22%)	6 (0.35%)	7 (0.30%)	4 (0.24%)	11 (0.27%)
Risperidone	461 (20.0%)	29 (1.71%)	345 (14.6%)	145 (8.80%)	490 (12.2%)
Haloperidol	586 (25.4%)	84 (4.95%)	414 (17.6%)	256 (15.5%)	670 (16.7%)
Quetiapine	1053 (45.6%)	103 (6.07%)	779 (33.1%)	377 (22.9%)	1156 (28.9%)
Sulpiride	262 (11.4%)	273 (16.1%)	333 (14.1%)	202 (12.3%)	535 (13.4%)
Levomepromazine	38 (1.65%)	7 (0.41%)	31 (1.32%)	14 (0.85%)	45 (1.12%)
Clozapine	20 (0.87%)	2 (0.12%)	17 (0.72%)	5 (0.30%)	22 (0.55%)
Tiapride	41 (1.78%)	10 (0.59%)	41 (1.74%)	10 (0.61%)	51 (1.27%)
Olanzapine	106 (4.59%)	14 (0.83%)	93 (3.95%)	27 (1.64%)	120 (3.00%)
Paliperidone	21 (0.91%)	6 (0.35%)	23 (0.98%)	4 (0.24%)	27 (0.67%)
Lithium	10 (0.43%)	0 (0.00%)	9 (0.38%)	1 (0.06%)	10 (0.25%)
Chlorpromazine	9 (0.39%)	0 (0.00%)	5 (0.21%)	4 (0.24%)	9 (0.22%)
Aripiprazole	38 (1.65%)	11 (0.65%)	40 (1.70%)	9 (0.55%)	49 (1.22%)

Supplementary Table 7. Comparison of patients' characteristics in training and test sets.

	Total N=4003	Training set N=3003	Test set N=1000	overall p
Psychotic disorder	2307 (57.6%)	1720 (57.3%)	587 (58.7%)	0.452
Depressive disorder	2356 (58.9%)	1752 (58.3%)	604 (60.4%)	0.268
Gender: female	2802 (70.0%)	2098 (69.9%)	704 (70.4%)	0.779
Age	85 [80;89]	85 [80;89]	85 [80;90]	0.441
Nursing home	943 (23.6%)	717 (23.9%)	226 (22.6%)	0.435
Hypertension	2398 (59.9%)	1811 (60.3%)	587 (58.7%)	0.390
Diabetes mellitus	1044 (26.1%)	798 (26.6%)	246 (24.6%)	0.234
Dyslipidemia	2059 (51.4%)	1534 (51.1%)	525 (52.5%)	0.459
Thyroid disease	764 (19.1%)	580 (19.3%)	184 (18.4%)	0.555
Parkinson's disease	193 (4.82%)	141 (4.70%)	52 (5.20%)	0.575
Stroke	1212 (30.3%)	904 (30.1%)	308 (30.8%)	0.707
Cardiovascular disease	790 (19.7%)	593 (19.7%)	197 (19.7%)	1.000
Traumatic brain injury	656 (16.4%)	490 (16.3%)	166 (16.6%)	0.873
Antipsychotic treatment	2095 (52.3%)	1562 (52.0%)	533 (53.3%)	0.504
No. of antipsychotic treats.	1 [1;2]	1 [1;2]	1 [1;2]	0.927
Changes from antipsy. to antidep.:				0.857
No changes	370 (17.7%)	274 (17.5%)	96 (18.0%)	
Some change	1725 (82.3%)	1288 (82.5%)	437 (82.0%)	
No. of changes from antipsy. to antidep.	1 [1;1]	1 [1;1]	1 [1;1]	0.612
Antidepressant treatment	2719 (67.9%)	2022 (67.3%)	697 (69.7%)	0.177
No. of antidepressant treats.	1 [1;2]	1 [1;2]	1 [1;2]	0.980
Changes from antidep. to antipsy.:				0.967
No changes	1463 (53.8%)	1087 (53.8%)	376 (53.9%)	
Some change	1256 (46.2%)	935 (46.2%)	321 (46.1%)	
No. of changes from antidep. to antipsy.	1 [1;1]	1 [1;1]	1 [1;1]	0.666
Sedation level:				0.578
None	1226 (30.6%)	920 (30.6%)	306 (30.6%)	
Minimum	215 (5.37%)	164 (5.46%)	51 (5.10%)	
Mild	1331 (33.3%)	1012 (33.7%)	319 (31.9%)	
Moderate	1209 (30.2%)	889 (29.6%)	320 (32.0%)	
Deep	22 (0.55%)	18 (0.60%)	4 (0.40%)	
Antipsychotic treatment	2095 (52.3%)	1562 (52.0%)	533 (53.3%)	0.504
Antidepressant treatment	2719 (67.9%)	2022 (67.3%)	697 (69.7%)	0.177
Trazodone	1519 (37.9%)	1131 (37.7%)	388 (38.8%)	0.545
Paroxetine	320 (7.99%)	241 (8.03%)	79 (7.90%)	0.953
Citalopram	274 (6.84%)	201 (6.69%)	73 (7.30%)	0.558
Mirtazapine	494 (12.3%)	375 (12.5%)	119 (11.9%)	0.664
Duloxetine	330 (8.24%)	242 (8.06%)	88 (8.80%)	0.502
Amitriptyline	180 (4.50%)	132 (4.40%)	48 (4.80%)	0.655
Desvenlafaxine	81 (2.02%)	54 (1.80%)	27 (2.70%)	0.104
Fluoxetine	108 (2.70%)	77 (2.56%)	31 (3.10%)	0.428

Sertraline	402 (10.0%)	294 (9.79%)	108 (10.8%)	0.390
Escitalopram	651 (16.3%)	473 (15.8%)	178 (17.8%)	0.141
Mianserin	58 (1.45%)	44 (1.47%)	14 (1.40%)	1.000
Venlafaxine	204 (5.10%)	161 (5.36%)	43 (4.30%)	0.215
Bupropion	37 (0.92%)	32 (1.07%)	5 (0.50%)	0.153
Vortioxetine	58 (1.45%)	42 (1.40%)	16 (1.60%)	0.757
Agomelatine	38 (0.95%)	28 (0.93%)	10 (1.00%)	0.998
Nortriptyline	2 (0.05%)	2 (0.07%)	0 (0.00%)	1.000
Maprotiline	4 (0.10%)	4 (0.13%)	0 (0.00%)	0.578
Clomipramine	13 (0.32%)	10 (0.33%)	3 (0.30%)	1.000
Fluvoxamine	11 (0.27%)	8 (0.27%)	3 (0.30%)	0.742
Risperidone	490 (12.2%)	375 (12.5%)	115 (11.5%)	0.442
Haloperidol	670 (16.7%)	505 (16.8%)	165 (16.5%)	0.855
Quetiapine	1156 (28.9%)	854 (28.4%)	302 (30.2%)	0.306
Sulpiride	535 (13.4%)	392 (13.1%)	143 (14.3%)	0.342
Levomepromazine	45 (1.12%)	34 (1.13%)	11 (1.10%)	1.000
Clozapine	22 (0.55%)	18 (0.60%)	4 (0.40%)	0.623
Tiapride	51 (1.27%)	37 (1.23%)	14 (1.40%)	0.805
Olanzapine	120 (3.00%)	85 (2.83%)	35 (3.50%)	0.333
Paliperidone	27 (0.67%)	25 (0.83%)	2 (0.20%)	0.058
Lithium	10 (0.25%)	7 (0.23%)	3 (0.30%)	0.718
Chlorpromazine	9 (0.22%)	8 (0.27%)	1 (0.10%)	0.466
Aripiprazole	49 (1.22%)	33 (1.10%)	16 (1.60%)	0.279

antidep., antidepressant; antipsy., antipsychotic

Supplementary Table 8. Relevance of variables for selected models.

Model predicting psychotic disorders		Model predicting depressive disorders	
Variable	MDA	Variable	MDA
Risperidone	29.63	No. of antidepressant treats.	46.11
Sedation level	29.55	Escitalopram	23.29
Quetiapine	24.68	Sedation level	18.41
Haloperidol	23.43	Sertraline	17.13
No. of antipsychotic treats.	22.88	Age	17.09
Sulpiride	16.44	Duloxetine	16.38
Trazodone	14.04	Citalopram	13.02
No. of changes from antidep. to antipsy.	12.67	No. of changes from antipsy. to antidep.	12.66
Parkinson's disease	12.38	Mirtazapine	11.8
Nursing home	9.37	Trazodone	7.7
Sertraline	8.25	Risperidone	6.75
Vortioxetine	8.09	Vortioxetine	6.43
Paroxetine	5.32	Venlafaxine	6.03
Mirtazapine	5	Paroxetine	5.4
Escitalopram	4	Desvenlafaxine	5.17
Olanzapine	3.06	Quetiapine	4.31
Stroke	2.67	Nursing home	4.09
Fluoxetine	2.2	Fluoxetine	3.83
Hypertension	1.8	Stroke	3.69
Tiapride	1.62	Olanzapine	3.09
Duloxetine	1.51	Hypertension	2.75
Citalopram	1.4	Diabetes mellitus	2.34
Gender	1.39	Sulpiride	2.33
Venlafaxine	1.36	Aripiprazole	2.3
Traumatic brain injury	0.9	Tiapride	1.97
Amitriptyline	0.78	Parkinson's disease	1.58
Cardiovascular disease	0.02	Thyroid disease	1.34
Desvenlafaxine	-0.06	Amitriptyline	1.12
Mianserin	-0.18	Gender	1.06
Dyslipidemia	-0.56	Levomepromazine	0.24
Levomepromazine	-1.1	Haloperidol	-0.13
Thyroid disease	-1.23	Cardiovascular disease	-0.35
Aripiprazole	-1.53	Traumatic brain injury	-1.99
Age	-1.71	Dyslipidemia	-2.33
Diabetes mellitus	-3.92	Mianserin	-3.53

MDA, mean decrease accuracy; antidep., antidepressant; antipsy., antipsychotic

REFERENCES

[1] Vilalta-Franch J, Lozano-Gallego M, Hernández-Ferrándiz M, Llinàs-Reglà J, López-Pousa S, López OL (1999) The Neuropsychiatric Inventory. Psychometric properties of its adaptation into Spanish. *Rev Neurol* **29**, 15-19.

[2] Alam MZ, Rahman MS, Rahman MS (2019) A random forest based predictor for medical data classification using feature ranking. *Inform Med Unlocked* **15**, 100180.

[3] Dai B, Chen R, Zhu S, Zhang W (2018) Using random forest algorithm for breast cancer diagnosis. *2018 International Symposium on Computer, Consumer and Control (IS3C)*, pp. 449–452.

[4] Fabris F, Magalhães JP, Freitas AA (2017) A review of supervised machine learning applied to ageing research. *Biogerontology* **18**, 171–188.

[5] Butt DA, Tu K, Young J, Green D, Wang M, Ivers N, Jaakkimainen L, Lam R, Guttman M (2014) A validation study of administrative data algorithms to identify patients with Parkinsonism with prevalence and incidence trends. *Neuroepidemiology* **43**, 28–37.

[6] Doupe P, Faghmous J, Basu S (2019) Machine learning for health services researchers. *Value Health* **22**, 808–815.

[7] Biau G (2012) Analysis of a random forests model. *J Mach Learn Res* **13**, 1063–1095.

[8] Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. *Biometrics* **33**, 159–174.