Received: 26 November 2021 | Accepted: 17 January 2023

DOI: 10.1111/jnp.12307

the british psychological society
promoting excellence in psychology

RESEARCH ARTICLE

One-year prediction of cognitive decline following cognitive-stimulation from real-world data

Borja Camino-Pontes[1] | **Francisco Gonzalez-Lopez**[2] | **Gonzalo Santamaría-Gomez**[2] | **Antonio Javier Sutil-Jimenez**[3] | **Carolina Sastre-Barrios**[2] | **Iñigo Fernandez de Pierola**[2] | **Jesus M. Cortes**[1,4,5]

[1]Biocruces-Bizkaia Health Research Institute, Barakaldo, Spain

[2]NeuronUP Labs, Logroño, Spain

[3]Department of Personality, Evaluation and Psychological Treatment, University of Granada, Granada, Spain

[4]IKERBASQUE: The Basque Foundation for Science, Bilbao, Spain

[5]Department of Cell Biology and Histology, University of the Basque Country, Leioa, Spain

Correspondence

Jesus M. Cortes, Biocruces-Bizkaia Health Research Institute, Cruces Plaza, 48903, Barakaldo, Bizkaia, Spain.
Email: jesus.m.cortes@gmail.com

Funding information

Agencia de Desarrollo Economico de La Rioja ADER, Grant/Award Number: 2019-I-CHE-00020; Centro para el Desarrollo Tecnológico Industrial, Grant/Award Number: EXP00117191/IDI-20190726; Research and development funds allocated by NeuronUP from its own resources.

Abstract

Clinical evidence based on real-world data (RWD) is accumulating exponentially providing larger sample sizes available, which demand novel methods to deal with the enhanced heterogeneity of the data. Here, we used RWD to assess the prediction of cognitive decline in a large heterogeneous sample of participants being enrolled with cognitive stimulation, a phenomenon that is of great interest to clinicians but that is riddled with difficulties and limitations. More precisely, from a multitude of neuropsychological Training Materials (TMs), we asked whether was possible to accurately predict an individual's cognitive decline one year after being tested. In particular, we performed longitudinal modelling of the scores obtained from 215 different tests, grouped into 29 cognitive domains, a total of 124,610 instances from 7902 participants (40% male, 46% female, 14% not indicated), each performing an average of 16 tests. Employing a machine learning approach based on ROC analysis and cross-validation techniques to overcome overfitting, we show that different TMs belonging to several cognitive domains can accurately predict cognitive decline, while other domains perform poorly, suggesting that the ability to predict decline one year later is not specific to any particular domain, but is rather widely distributed across domains. Moreover, when addressing the same problem between individuals with a common diagnosed label, we found that some domains had more accurate classification for conditions such as Parkinson's disease and Down syndrome, whereas they are less accurate for Alzheimer's disease or multiple sclerosis. Future research should combine

similar approaches to ours with standard neuropsychological measurements to enhance interpretability and the possibility of generalizing across different cohorts.

KEYWORDS
cognitive decline, longitudinal data, neuropsychology, prediction modelling, real world data

INTRODUCTION

Recent technological advances have expanded the possibilities of collecting and storing large amounts of data, paving the way for the more extended use of Real-World Data (RWD) or Real- World Evidence (RWE) in clinical research. Food and Drug Administration (FDA) declares that Real-World Data (RWD) are data relating to patient health status and/or the delivery of health care routinely collected from variety of sources (U.S. Food & Drug Administration, 2022). Sherman et al. (2016) also emphasized that RWD refers to data collected in clinical, home or community settings and not in academic environments. The combination with Electronic Health Records (EHRs) and other sources it makes RWD to open novel possibilities in clinical research (Jarow et al., 2017; Sherman et al., 2016). Moreover, as the physical devices used to collect such data are increasingly portable, information can be obtained from multiple locations at the same time, increasing the sample size of the research study in question (Milne-Ives et al., 2020). The combination of EHRs, portable devices and modelling through techniques like machine learning is a growing area in clinical research and neuropsychology (Mawdsley et al., 2021; Vaccaro et al., 2021). Following this line, this study focuses on the use of RWD to assess cognitive decline in a large sample of participants receiving a cognitive stimulation program, a challenge that must overcome several difficulties and limitations (Graham et al., 2020).

Despite numerous efforts to understand the cognitive impairment associated with different diseases and conditions, for a review see (Solomon & Soininen, 2015), still there is room to improve previous models. Strict inclusion and exclusion criteria often limit the generalizability of the findings to different cohorts. EHRs and RWD can help overcome these limitations (Jaakkimainen et al., 2016; Ponjoan et al., 2019, 2020; Shehzad et al., 2020), and the use of machine learning seems suitable to model such heterogeneous data (Seccia et al., 2021). Indeed, many attempts during the last years have advanced the detection of dementia using RWD (Ben Miled et al., 2020; Kim et al., 2017; Luo et al., 2020; Nori et al., 2019; Wolfsgruber et al., 2014). Recommendations based on the use of RWD to study cognitive decline have shown the benefits of therapy to delay the severity of dementia-related dependencies with institutionalization (Davis et al., 2018), highlighting the relevance of cognitive stimulation and neurorehabilitation to delay dementia and cognitive decline. The use of online platforms for interventions by clinical neuropsychologists and for rehabilitation programs is becoming popular (Irazoki et al., 2020; Pertíñez & Linares, 2015), yet despite evidence of the efficacy of such platforms (Arroyo-Ferrer et al., 2021; Mendoza Laiz et al., 2018) and their utility in predicting cognitive decline is more limited.

Most research into predictive models for dementia have been based on a combination of biomarkers and clinical measurements, and on the relevance of those parameters that predict conversion from cognitive impairment to Alzheimer's disease (AD) (Bucholc et al., 2019; Casanova et al., 2020; Dubois et al., 2018; Ewers et al., 2012; Ezzati et al., 2019; Lee et al., 2019; Li et al., 2013; Li & Fan, 2019; Palmqvist et al., 2012; Young et al., 2014). Other studies have combined imaging with biomarkers (Bucholc et al., 2019; Ewers et al., 2012; Gleason et al., 2018; Gomar et al., 2014; Lee et al., 2019; Pereira et al., 2018; Tabarestani et al., 2020; Tatsuoka et al., 2013) or retinal morphology (Murueta-Goyena et al., 2019, 2021), while cognitive tests have also been used to predict decline in elderly participants, often producing results as good as those from biomarkers studies (Bucholc et al., 2019; Fields et al., 2011; Gleason et al., 2018;

Gomar et al., 2014; Lee et al., 2006, 2019; Li et al., 2013). In relation to the use of RWD to predict dementia, lifestyle has shown promising for predicting the risk of dementia (Fouladvand et al., 2019; Ritchie et al., 2018; Sindi et al., 2015) and indeed, in multiple sclerosis (MS) cognitive decline has been assessed using RWD as a precursor of dementia to evaluate the prognosis and evolution of this condition (Cohen et al., 2020; Matthews et al., 2020).

Here, we hypothesize that by adopting a machine learning approach based on RWD obtained from an online cognitive stimulation platform, it might be possible to predict cognitive decline one year after testing from the scores obtained in different training materials (TMs). If this were the case, it would pave the way to identify patients potentially susceptible to cognitive decline, allowing patient-specific programs to be designed to slow this decline.

MATERIALS AND METHODS

Data

All the participants studied in this work were enrolled in a subject-specific cognitive stimulation program using NeuronUP, a cloud platform for cognitive stimulation that contains 215 different TMs grouped into 29 different cognitive domains and with different number of TMs completed per domain (Table 1). For this study, 203 different TMs were considered for the analyses, as 12 of them did not satisfy at least one of the items in the TM selection criteria (see below). Analyses were performed on 124,610 samples from 7902 participants (40% male, 46% female, 14% not indicated), each performing an average of 16 TMs. The average age of the participants was 50 years of age (std. dev. = 24).[1] The patients participating in this study have carried out the TMs in multiple countries at home, private clinics or stimulation centres, following a patient-specific simulation program designed by their clinical responsible. Table 2 shows age statistics of the studied samples. It is important to emphasize that a precise diagnosis of the different participants was not performed, but this label (cf. Table 2, Alzheimer's disease, Parkinson's disease, multiple sclerosis or Down's syndrome) was assigned in relation to the stimulation centre of origin, since those centres are specialized in those conditions. The label *Other* in Table 2 referred to participants who were belonging to any of those centres and had unknown diagnostic label.

Ethical considerations

This study was exempted from an Ethics Review Panel as it makes use of low-risk retrospective research from existing collections of data that contain non-identifiable data, and where all participants provided information before collecting the data.

In relation to the signed consent, all NeuronUP stimulation centres signed strict agreements for the use of data for research purposes. The agreement establishes that all responsible centres have, before collecting the data, to adequately inform on: (1) That the information collected will be used for various purposes related to research in prevention, diagnosis and treatment of diseases, and the promotion of health in society. (2) That there is the possibility of asking everything that the participant needs or does not understand. (3) That participation is completely voluntary and that at any time the participant can withdraw without giving any explanation. In this case, the participant's data will be removed from the NeuronUP platform. (4) That the NeuronUP data team will never have access to the participant's personal data, and that their identity will be encrypted with an alphanumeric code and will not contain any personal or identifying data of the participant. (5) That the participant's data will never be transferred

[1]Note that the age reported could sometimes vary with respect to chronological age, due to some inaccuracies in its collection, *i.e.*: some participants did not report it, others gave it incorrectly or it was sometimes provided by third parties (monitor, clinician or family members).

TABLE 1 Cognitive domains and statistics of instances and training materials (TMs).

Cognitive domain name	Number of TMs per domain	Statistics of instances per domain at T_0 (mean ± std. dev)
Alternating Attention	5	7.5 ± 10.3
Auditory Gnosis	1	19.6 ± 17.9
Body Schema	2	7.0 ± 7.1
Comprehension	3	20.0 ± 23.8
Cooking and Cleaning	2	12.6 ± 12.2
Decision Making (EF)	3	10.8 ± 11.4
Discrimination	4	27.8 ± 34.5
Episodic Memory	16	14.4 ± 16.3
Expression	1	19.9 ± 17.0
Flexibility (EF)	4	11.2 ± 12.8
Hemineglect	1	12.1 ± 8.8
Inhibition (EF)	1	9.2 ± 11.5
Naming	1	12.8 ± 16.6
Place (Orientation)	2	19.7 ± 28.7
Planning (EF)	11	12.3 ± 13.7
Processing Speed	21	20.6 ± 25.8
Reasoning (EF)	10	6.4 ± 9.9
Selective Attention	22	10.7 ± 15.9
Semantic Memory	11	7.2 ± 16.2
Social Cognition	6	9.5 ± 12.4
Spatial Relation	10	27.5 ± 29.5
Spatial Visualization	2	13.8 ± 17.1
Sustained Attention	10	10.9 ± 12.6
Time (Orientation)	1	8.1 ± 10.1
Time Estimation (EF)	2	15.4 ± 20.4
Visoconstructive Praxis	1	9.1 ± 8.1
Visual Gnosis	11	13.4 ± 18.7
Vocabulary	11	19.4 ± 26.8
Working Memory (EF)	22	14.6 ± 22.4

TABLE 2 Size and age distribution across different diagnostic labels.

Diagnosis	#Participants	Mean age[b]	Std. dev. Age[b]
Alzheimer's disease	1157	76.35	15.12
Down's disease	744	33.65	8.81
Multiple sclerosis	653	56.54	15.66
Parkinson's disease	163	71.26	7.07
Other[a]	5185	44.07	23.79

[a]Participants with unknown diagnosis.

[b]Calculated over the number of subjects with collected age.

to third parties or companies, although the possibility of NeuronUP collaborating with other partners in research projects is contemplated. All data used in this study come from participants who have given their consent by signing an electronic form.

Instances, TM selection criteria, prediction model and data-driven definition of cognitive decline

After a participant completed a TM, three different variables were measured: *Trues*, the number of correct answers; *Time* or duration, required for completing the TM; and *Level*, an internal difficulty measure created for each TM by the NeuronUP Neuropsychology team. For example, for the TM called 'Additions', there are 5 difficulty levels, basic (B), easy (E), medium (M), difficult (D) and advanced (A), each one differentiated by the number of addends, the number of addend digits, and whether or not there are carry-overs in the sum. The basic level corresponds to 2 addends, only 1 digit and no carry-overs, and at the opposite extreme, the advanced level corresponds to 4 addends, 6 digits and the possibility of carry-overs. For the rest of the TMs, the level of difficulty has been developed specifically for each TM. From here on, these three features (Trues, Time and Level) were transformed into percentiles calculated from all subject's performances who performed the TM within the NeuronUP platform. The triplet of percentile values $(p_{Trues}, p_{Time}, p_{Level})$ defined an instance, which corresponded to each completion of a given TM, and then were used for prediction by the machine-learning analyses. It is important to emphasize that working with percentile values is relevant to modelling, since percentiles, as opposed to raw data, provide comparative information on the subject's performance relative to others in the sample.

The TM selection criteria consisted of: 1. A given participant had to complete the same TM at the baseline time point (T_0) and 1 year later (T_1); 2. For each TM, the minimum sample size was 50 instances; 3. For each TM, at least 5 instances belonged to the class of cognitive decline (see below for details).

For the prediction model, final measures for T_0 and T_1 of the variables 'Trues', 'Time' and 'Level' were obtained. In particular, for the T_0 (the 'basal set'), a time window of 15 days was considered to collect the different values from which the average across these days was calculated as the final value for the machine learning analysis. For the T_1, a similar strategy was applied but using a 90-day time window centred one year after the mean date of the basal instance. We also introduced the constraint of having an equal number of T_1 instances than the one in the T_0 set, allowing equivalent statistics for both the T_0 and T_1 set of instances. Thus, if we selected 5 instances in the basal set and 10 instances existed within the 365 ± 90 day window, we finally chose the 5 instances out of those 10 that were closest to the time point of +365 days. Our aim was to predict from the performance at T_0 the classified event at T_1.

Before performing any prediction, we needed to identify and label the participants who had cognitive decline. For this we used the NeuronUP Score (represented by s), a different one for each instance, and calculated based on participant's performance by using a formula that combines *Trues*, *Times*, and *Level* and returns an index ranging between 0 and 100. This score is a novel quantitative index created to simplify a participant's performance and to facilitate the comparisons in the follow-up, allowing longitudinal data of the same participant to be modelled while obtaining precise trajectories of individual performance. In this modelization, we obtained one score at baseline (s_0) and another one, 1 year after (s_1). We then calculated the difference ($diff = s_1 - s_0$) which allowed defining 'cognitive decline' (class 1) after fixing a threshold (Z_{th}) in the Z-score of the variable *diff*. Other instances were defined as class 0, i.e.: not having cognitive decline during the year after baseline. We have used several choices for the threshold, namely, $Z_{th} = -.5$, -1.0, -1.5, but for most of the analyses we referred in this study we selected $Z_{th} = -1$. This choice, corresponding to Z values smaller than the mean minus one sigma,[2] was made satisfying two conditions simultaneously, that the tail of the distribution to the left of the threshold be the smallest possible, and also that the imbalance between class 1 and class 0 in all cases was not very pronounced. It is important to emphasize that NeuronUP score (s) were only used to calculate the labels class 1 and class 0. Once the corresponding label was assigned to each of the instances, all the predictive models (one per TM) were trained based on the subsample of participants who performed the corresponding TM from their triplet values of $p_{Trues}, p_{Times}, p_{Level}$.

[2]Straight to obtain from the definition $Z_{diff} = (diff - \mu_{diff})/\sigma_{diff}$, with μ_{diff} and σ_{diff} equal to the mean and standard deviation of all values of variable $diff = s_1 - s_0$ across all subjects participating in the same TM.

Machine learning classification

A machine learning analysis was performed to predict the occurrence of class 1 events, i.e.: participants who experienced decline 1 year after the NeuronUP stimulation session at T_0. The analyses were performed using the *scikit-learn* package [https://www.scikit-learn.org] running in Python 3 [https://www.python.org/]. In particular, we performed an analysis of ROC curves using the *LogisticRegression* function with *L2* regularization. Other classifiers such as random forest, support vector machine, and decision trees were also used but their performance was systematically equal or lower than the one achieved by logistic regression, which is the case we report here. *L2* regularization was justified because our logistic regression model only had three input variables p_{Trues}, p_{Times}, p_{Level} and therefore other strategies such as *L1* regularization were not needed (the latter well known to work out for large number of model variables and the capability to nullify some of them). Due to the unbalanced nature of our data (most of the participants had no cognitive decline), it was necessary to correct this bias by adjusting the weights of the classes with an inversely proportional factor (F_i) that affects each class, i.e.:

$$F_i = \frac{Total\ number\ of\ samples}{Samples\ in\ class_i * Number\ of\ classes} \quad (1)$$

Moreover, to overcome overfitting, a 5-fold cross-validation strategy was employed to assess the generalization of the model. Specifically, for the logistic regression classifier the predicted probabilities P were obtained by:

$$P = \frac{1}{[1 + e^{-y}]} \quad (2)$$

where y was obtained by fitting the logistic regression model with input variables equal to p_{Trues}, p_{Times}, p_{Level}. Finally, ROC curves were obtained using the *roc_curve* function. To determine the optimal cut-off value, we maximized both sensitivity and specificity by using a geometrical argument that consists in choosing as the cut-off value, the point belonging to the ROC curve closest to the point (0,1) in the representation (sensitivity) versus (1 - specificity). This choice maximizes simultaneously (and separately) the two metrics sensitivity and specificity. Other criteria can be used such as the maximization of the Youden Index, defined as sensitivity + specificity - 1, thus maximizing the sum of the two metrics sensitivity and specificity. Note that the determination of the optimal cut-off value θ^{opt} is clinically relevant as it simplifies the strategy for classification. In particular, given the performance of a new participant represented by the triplet of variables p^{new}_{Trues}, p^{new}_{Time}, p^{new}_{Level}, the classification consists in evaluating the logistic regression output and see whether it is below or above the cut-off value, and therefore, establishing the belonging to class 0 or class 1, respectively. For quantification of classification goodness, we chose the area under the curve (AUC) as the metric of the model's performance. The higher the AUC, the better the performance of the model at discriminating between class 1 and class 0 instances. Our classification strategy has been performed individually for each TM. Other multivariate classification strategies, for instance by combining scores of different TMs within the same cognitive domain, are difficult to assess in our RWD situation as not all the participants completed the same set of TMs.

RESULTS

A total number of 203 predictions models were built, each one based on a different TM existing in the NeuronUP cognitive stimulation platform. We aimed predictions of cognitive decline after 12 months based on the basal performance (Figure 1). Data analysed consisted in 124,610 samples from 7902 participants, each performing 16 tests on average. Three major steps were followed; first, an average basal score at time T_0 (s_0) was collected (Figure 1a), and 12 months later, another one at T_1 (s_1). Second, these two scores were used for data labelling, defining the variable $diff = s_1 - s_0$ and assessing whether its Z-score was

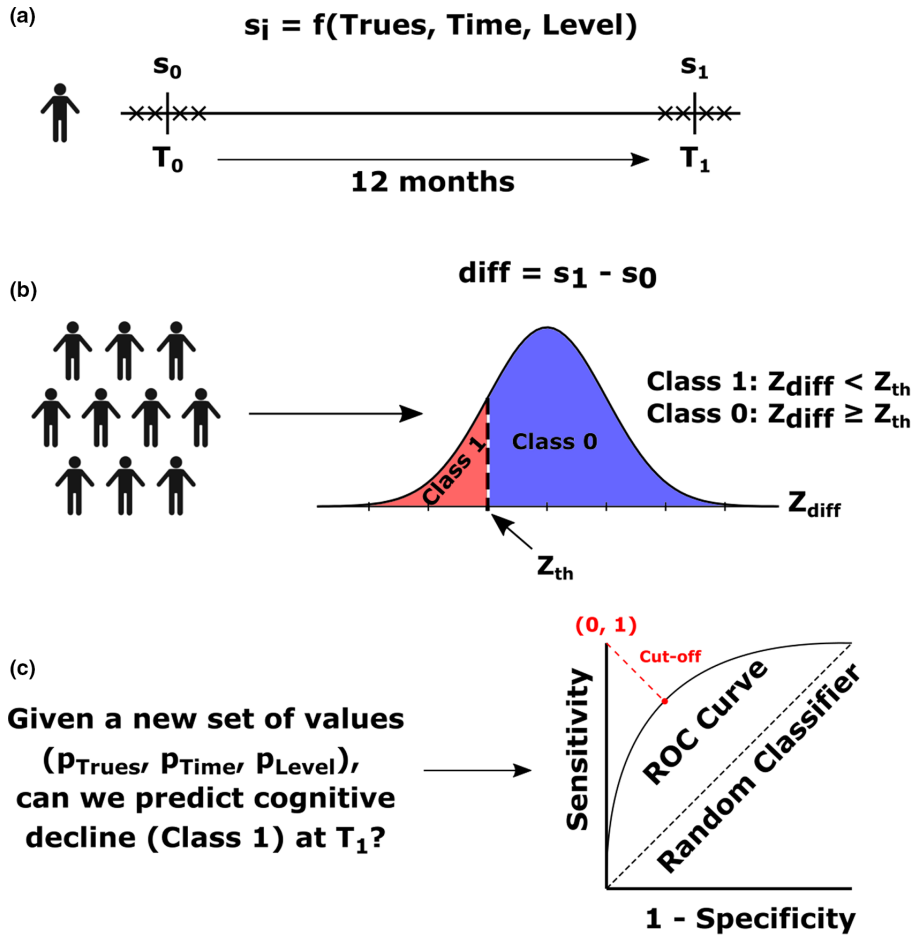

FIGURE 1 Prediction of cognitive decline using Real-World-Data. For this study, 203 different classification models, each one for a different TM, were built from a total of 124,000 samples obtained from 7902 different participants who performed each an average of 16 TMs. The prediction of cognitive decline was performed using the following strategy. a, for each participant we identified different scores (si) at the baseline, T_0, that were finally averaged into s_0. Following-up the given participant, we identified a set of scores at time $T_1 = T_0 + 12$ months, and then averaged into s_1. b, both s_0 and s_1 were then used for data labelling by defining $diff = s_1 - s_0$, calculating the set of Z-scores and assigning the label of cognitive decline, Class 1, for the Z values smaller or equal than Z_{th}, and Class 0 otherwise. c, after labelling all the instances, we trained and tested different models for the prediction of Class 1, using as input variables the percentiles values (p) of *Trues* (the number of correct answers), *Time* (required to complete the TM), and *Level* (an internal difficulty measure for each TM). ROC analyses provided cut-off values maximizing both sensitivity and specificity simultaneously.

smaller than a threshold (Z_{th}) to assign that participant to class 1 (*cognitive decline*) or to the class 0 elsewhere (*no cognitive decline*) (Figure 1b). Next, we built classification models for each TM, using the set of instances p_{Trues}, p_{Time}, p_{Level} at T_0 from all the participants who completed the corresponding TM (Figure 1c). Because our prediction models were based on ROC analysis, for each TM we determined a cut-off value (marked in red in Figure 1c).

For the determination of the Z_{th}, we tried to satisfy two conditions simultaneously, that the tail of the distribution to the left of the threshold be the smallest possible, and that the imbalance between class 1 and class 0 was not very pronounced. Results of classification accuracy as measured by AUC for $Z_{th} = -.5, -1.0$, and -1.5 are shown in Figure S1-S2. In general, the higher the threshold, the higher the AUC, but this is driven by an increase in the imbalance between class 1 and class 0 events. Therefore, although in our classification method we compensated for the imbalance between the classes (see Meth-

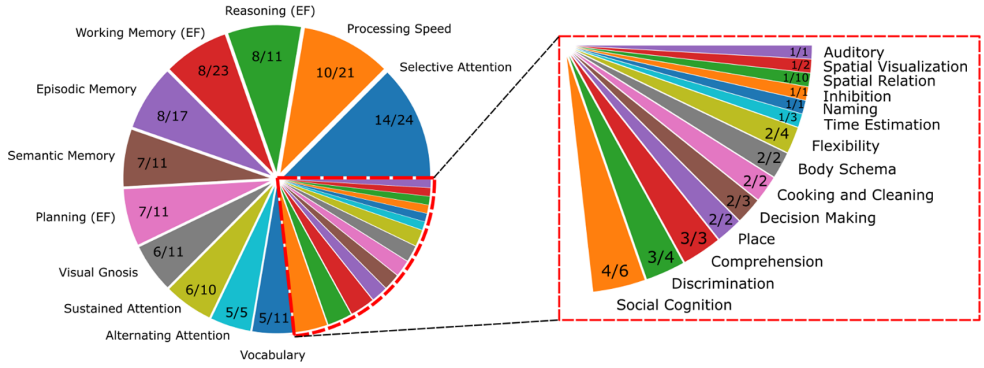

FIGURE 2 Number of tests per cognitive domain that had high accuracy (AUC > .75) to predict cognitive decline. One slice in the pie chart represents one cognitive domain. The first number within each slice indicates how many of the TMs within that domain provided an AUC > .75, while the second corresponds to the total number of TMs within that domain. The size of each slice scales up with the number of TMs with an AUC > .75.

ods), for extreme imbalances this correction was not sufficient. In particular, the imbalance proportions for the respective thresholds −.5, −1.0, −1.5, were 39%, 12%, and 5%. Based on the results, we finally chose $Z_{th} = -1.0$ for further analyses, as the 12% imbalance allowed us to have enough statistics while keeping a good compromise for the definition of cognitive decline (the smaller the threshold, the better defined).

Our first results showed that not all the domains available were very accurate in predicting cognitive decline, emphasizing that some domains were more useful than others to predict decline 1 year after TM completion. In particular, of the 29 different domains assessed in this study (shown in Table 1), four of them did not contain any TM that satisfied an AUC > .75. These *moderate classified* domains were Expression, Hemineglect, Time Orientation and Visoconstructive Praxis. For the remaining 25 domains (shown in Figure 2), at least one of the TM achieved a level of classification with an AUC > .75. The number of the TM that provided AUC > .75 is also shown in the same figure.

For the domains containing the larger number of TMs with AUC > .75, we next chose the TM with highest AUC and represented their ROC curves (Figure 3). Green ROC curves corresponded to the results from the training dataset (i.e.: the entire dataset), while the blue curves corresponding for cross-validated results. For each TM, individual cut-off value is also depicted and marked by a green arrow. Very remarkable domains such as Processing Speed, Selective Attention and Alternating Attention provided performances of AUC > .90.

Finally, we asked whether some domains were better predictors than others across different diagnostic labels (see Methods), assuming that some of them might have more severe deficits in specific cognitive areas than others and that those domains might therefore have greater predictive power. We followed the same methodology but we trained the different models within distinct specialized clinical institutions, which guaranteed that the participants from those institutions had a common diagnosed condition. In particular, we focus here on Alzheimer's disease, Parkinson's disease, multiple sclerosis and Down's syndrome, although for illustrative purposes we also provide results for the condition *Other*, meaning that those participants did not come from any of those institutions and had unknown diagnosis (see Table 2 for sample size and age distribution for each diagnostic label). For each of these diagnosed conditions, we calculated the percentage of correctly classified instances (PCCI), defined as the sum of true positives and true negatives divided by the total number of instances and across all TMs within a cognitive domain (each represented in a row in Table 3). Note that when comparing PCCIs across diagnostic labels, no domain had PCCIs > .85 for Alzheimer's disease, multiple sclerosis, or the *Other* condition. However, Parkinson's disease had a PCCI > .85 for Selective Attention, Visual Gnosis, Reasoning (EF), Alternating Attention and Sustained Attention. Similarly, Processing Speed, Visual Gnosis, Planning, Reasoning, Vocabulary and Alternating Attention provided TMs with the PCCI > .85 for Down's syndrome.

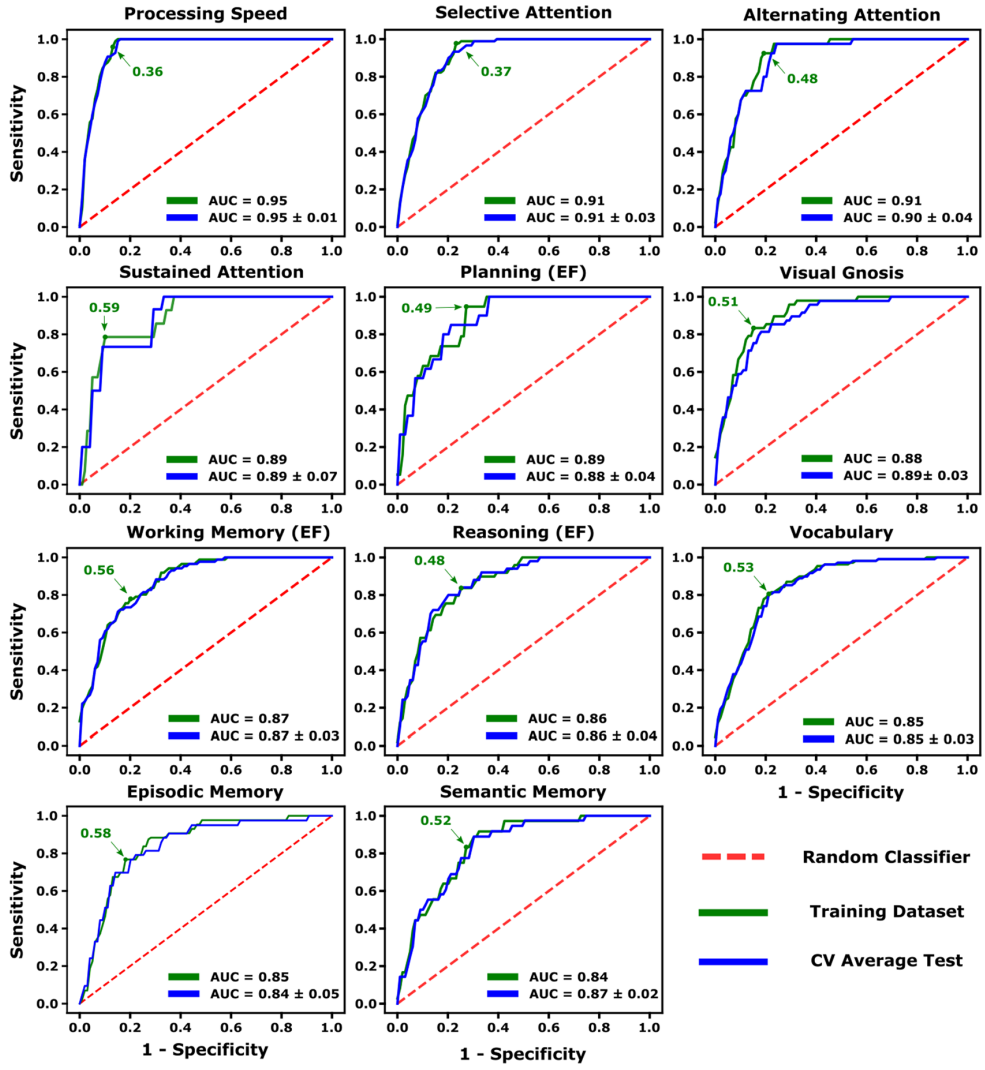

FIGURE 3 ROC curves for TM within domain with the best classification performance. Within each cognitive domain (as specified in each plot), we show the ROC curve of the most accurate TM based on AUC (the higher value, the better classification performance). The green curves correspond to the ROC curves obtained from the training dataset (i.e.: the entire dataset), while the blue curves show the mean value of the 5 different cross-validated (CV) ROC curves. Green arrows indicate the cut-off value for each TM. We also provided the ROC curve for the random classifier (red dashed line) for illustrative purposes only. The chosen domains in this figure correspond to those in Figure 2 that had larger numbers of TMs per domain with AUC > .75.

DISCUSSION

A combination of machine learning and RWD has been used to predict cognitive decline after one year in a large sample of participants receiving cognitive stimulation from different TMs. A total of 124,610 instances from 7902 different participants were used, each participant performing an average of 16 instances. As a result, we explored the predictive capacity of 203 different TMs grouped into 29 different cognitive domains. Overall, we have found good predictability for a wide range of domains, since 25 of the 29 domains analysed provided accuracy of AUC above .75. This is consistent with previous studies predicting cognitive decline from RWD and the risk of dementia, showing that there exist multiple different domains producing accurate predictions. For instance, a follow-up study of patients

TABLE 3 Percentage of correctly classified instances (PCCI) for each cognitive domain across different diagnoses.

	Alzheimer's disease	Parkinson's disease	Multiple sclerosis	Down's syndrome	Other
Processing speed	.76	.79	.64	**.86**	.76
Selective attention	.76	**.95**	.55	.83	.72
Visual gnosis	.67	**.87**	.52	**.85**	.73
Planning (EF)	.74	.79	.59	**.88**	.72
Working memory (EF)	.72	.69	.56	.71	.71
Reasoning (EF)	.74	**.88**	.57	**.87**	.76
Episodic memory	.78	.82	.53	.83	.73
Vocabulary	.77	.84	.54	**.85**	.79
semantic memory	.70	.74	.50	.84	.69
Sustained attention	.75	**.89**	.50	.82	.72
Alternating attention	.82	**.85**	.73	**.96**	.75

Note: The PCCI, defined as the sum of true positives and true negatives divided by the total number of instances, and calculated over all TMs belonging to each cognitive domain (name in rows). The chosen domains in this table correspond to those in Figure 2 that had larger numbers of TMs per domain. The percentage of incorrect responses can be simply calculated as 1 minus the values provided in this Table. PCCI values larger than .85 are shown in bold. The *Other* condition shows the results of participants who had unknown diagnosis.

with mild cognitive impairment showed that memory and executive function were the best predictors for deterioration, and these two domains continued to provide long-term accuracy (Li et al., 2013). In another study, the Trail Making Tests, assessing flexibility, inhibition and attention, showed good performance in predicting cognitive decline (Parikh et al., 2014), as also occurred during the preclinical stages of AD (Albert et al., 2001; Ewers et al., 2012). Verbal declarative memory was also shown to be a good predictor of rapid cognitive decline and disease progression (Sala et al., 2017), as reported previously (Grande et al., 2018). The clock drawing test, a common screening test for early cognitive impairment, was also shown to be a good predictor of AD (Gomar et al., 2014; Palmqvist et al., 2012). Likewise, episodic memory was highlighted as a good predictor using different tests (Gomar et al., 2014; Pereira et al., 2018; Tatsuoka et al., 2013), and along with executive function, both were seen to be good predictors in a three-year follow-up study (García-Herranz et al., 2016). Finally, and highlighting the fact that multiple domains can predict cognitive decline, some studies have shown the benefits of using global cognitive measures, defined as composites of multiple domains, to predict cognitive decline or conversion from mild to late cognitive impairment (Gavett et al., 2010; Gleason et al., 2018; Lee et al., 2006).

Our RWD prediction analyses to predict cognitive decline was also applied to subgroups of participants obtained from specialized diagnosis-focused institutions, thereby ensuring that the participants belonging to each institution had a common diagnosed condition. By measuring the PCCI[3] within each pathological condition, we found that no cognitive domains had a PCCI > .85 among patients diagnosed with Alzheimer's disease (AD) or for participants with Multiple Sclerosis (MS). By contrast, Parkinson's disease (PD) patients had a PCCI > .85 in the domains of Selective Attention, Visual Gnosis, Reasoning (EF), Sustained Attention, and Alternating Attention. Similarly, for Down's syndrome (DS) our analyses showed PCCI > .85 for the domains of Processing Speed, Visual Gnosis, Planning, Reasoning, Vocabulary and Alternating Attention. It is noteworthy that the Visual Gnosis domain performed well in both DS and PD. Thus, our RWD results showed that some cognitive domains are more specific than others to correctly predict cognitive decline in PD and DS patients but not in those diagnosed with AD and MS. Attention has been seen to be good predictors of decline in PD patients (Baiano et al., 2020; Pedersen et al., 2013) in full agreement with our results (Selective Attention PCCI = .95, Sustained Attention PCCI = .89, Alternating Attention PCCI = .85). Moreover, executive function was shown good predictors of cognitive decline in a longitudinal study, also consistent with our data (Reasoning PCCI = .88,

[3]Percentage of Correctly Classified Instances (*cf.* Methods).

Alternating Attention PCCI = .85). For DS, we found Planning (PCCI = .88), Reasoning (PCCI = .87), and Processing Speed (PCCI = .86), Vocabulary (PCCI = .85) and Alternating Attention (PCCI = .96) to have high specificity for cognitive decline, which might be related to the focusing of the neurocognitive profile on language and tasks (Næss et al., 2011; Silverman, 2007).

It is important to note that we chose here a supervised classification scheme to predict the class of cognitive decline after one year of being tested, but other strategies such as regression might also be used. The two strategies make possible to associate the values of the percentiles at T_0 with the functioning at T_1 (Figure S2), although a systematic regression study (for each test, domain and diagnostic groups) should be investigated in future research. The two strategies, moreover, could serve to precisely track the long-term trajectories of cognitive impairment with an adequate longitudinal follow-up of participants, overcoming current limitations in longitudinal studies due to small temporal resolution in these trajectories (to give some startling numbers, for some of the participants, the number of time points collected was greater than a thousand over a 6-year follow-up period; other participants performed the same TM 98% of the days in a period longer than 2 years). As a result, the RWD paradigm may be better than non-RWD approaches to characterize these trajectories, the latter typically involving strong methodological constraints, such as the same number of measurements or time points, and requiring equivalent time intervals to be used across subjects, and where the different variables to study are much more controlled in statistical sense. RWD offers multiple possibilities for clinicians and in particular, the ability to identify subjects susceptible to accelerated decline might help design subject-specific stimulation programs. Future studies assessing prediction of outcome continuously within such well-resolved time trajectories might be of high interest.

We are not suggesting, however that RWD works always better than traditional studies such as random control trials (RCTs). In fact, we conceive RWD and RCT as two complementary approaches. While RCT is able to maximize the internal validity (referring to control of variables ensuring that the independent variables cause the changes in the dependent variable) by using a restricted sample controlling confounding variables, RCT fails when trying to generalize and extrapolate the obtained results to clinical routine, and this affects all medical disciplines (Monti et al., 2018; Yuan et al., 2018). RWD, in combination with RCT, offers the possibility to maximize the external validity (limited in RCTs and referring to the possibility to generalize the results), by using a much larger sample and technologies to collect data in real environments (Monti et al., 2018; Yuan et al., 2018). Continuous monitoring and analysis of RWD could help bridge the efficacy–effectiveness gap enhancing long-term drug efficacy or interventions when applied to *imperfect* patients. Therefore, our study offers an advanced methodological analysis, which increases the data reliability and could also be tested in clinical trials.

Our study is not exempt from limitations, almost all of them related to the highly heterogeneous nature of RWD. First of all, since scores are not available for all TMs and all subjects, it is not feasible to apply well-known techniques, such as Principal Component Analysis, to the original data to obtain a dimensionality reduction across different TMs. Second, it is important to remark that we cannot control the effect that participation in the NeuronUP program has on the cognitive decline one year later, since all the participants were enrolled in the stimulation program. Although this is beyond the current scope, future studies could separate an intervened group versus non-intervened and assess the differences between the two patterns of decline. Third, there is a significant age difference between the *Other* and Down syndrome groups compared with participants with AD and PD (the latter groups being much older) and, to a lesser extent, with MS. Therefore, there exist a confounding factor of age in our study affecting the neurodegenerative groups (AD and PD), which requires additional strategies to overcome this effect when comparing performances between the groups. Forth, RWD is also heterogeneous in the number of TMs within each cognitive domain, and future studies are needed for assessing the internal consistency of each domain. A preliminary analysis as shown in Table 4 has shown that some domains have higher AUC consistency than others,[4] with the important limitation that the different TMs in each domain are not performed by the same group of participants, so we cannot control for this source of

[4]Consistency defined here simply by the inverse of the variance of all AUCs (each achieved by a different TM) within a given domain.

TABLE 4 AUC consistency of the prediction across TMs belonging to the same cognitive domain.

Cognitive domain	AUC consistency
Flexibility (EF)	120.63
Planning (EF)	135.64
Spatial visualization	143.57
Processing speed	145.28
Selective attention	158.55
Episodic memory	164.02
Sustained attention	240.14
Visual gnosis	252.80
Vocabulary	259.33
Reasoning (EF)	277.68
Working memory (EF)	299.44
Semantic memory	315.48
Time estimation (EF)	392.87
Social cognition	423.17
Decision making (EF)	426.48
Alternating attention	513.29
Cooking and cleaning	535.65
Spatial relation	538.49
Body schema	1225.38
Discrimination	1654.01
Comprehension	1930.93
Place (Orientation)	2910.17

Note: Consistency has been defined as 1/variance of AUC, each achieved by a different TM, and the variance calculated over all TMs within the same cognitive domain.

variability in our dataset. Fifth, the pathological conditions studied here are difficult to diagnose in their early stages (e.g., AD or PD) and thus, the labelling of these pathological conditions might sometimes be inaccurate. In fact, early-stage diagnoses of neurodegenerative diseases are still a challenge in clinical practice in part due to the similarities in some of the symptoms that define these conditions in their initial stages. In addition, as we have explained, the diagnoses used here have not been clinically obtained, but assigned according to the participant's institution, i.e. there are centres specialized in specific pathologies and we have assigned those diagnostic labels to the participants from those centres. Sixth, it could be thought that because the variable that defines the decline is diff $= s_1 - s_0$, and because the classification is performed using the variables p_{Trues}, p_{Time}, p_{Level}, which in turn are used for the calculation of s_0, there may be some circularity in the data (Lega et al., 2022; Oldham, 1962), which introduces a positive bias for the prediction. On the other hand, the logistic regression model was adjusted differently for each MT, maximizing the precision in the classification, and therefore with a different dependence on the input variables to the one used for the calculation of s_0. We have verified that, in general, the use of s_1 versus s_1-s_0 does not introduce such effects, being able to independently obtain a better or worse classification in each of the cases, thus confirming *a priori* a certain level of independence. Another limitation is related to the assignment between TMs and cognitive domains. The NeuronUP Neuropsychology team performed this task in our study, however future validation of such assignments using more standardized tools will be of interest to generalize our results to other cohorts. It is also very important to emphasize that the term 'Cognitive Decline' used in this study (i.e., data-driven defined declines one year after being tested) is not equivalent to 'Cognitive Impairment', a very well-defined condition at the early stages of neurodegenerative diseases and encompassing multiple clinical and neuropsychological deficits.

In fact, it would be possible that some of the participants in this study could have cognitive impairment in combination to cognitive decline. To fully correspond the association between cognitive decline and cognitive impairment further studies combining comprehensive cognitive and neuropsychological participant's evaluations together with ML methodologies are needed.

To conclude, by employing a machine learning methodology based on ROC analysis and cross-validation techniques to overcome overfitting, we show here that it is possible to predict cognitive decline one-year after testing when using RWD from cognitive stimulation. When applied to our entire sample without stratification, Processing Speed (AUC = .95), Selective Attention (AUC =.91), Alternating Attention (AUC = .91), Sustained Attention (AUC = .89), and Planning (AUC = .89) are the domains that most accurately predict cognitive decline. Moreover, we found that other combinations of domains performed well in specific PD and DS participants, although no domains stood out in predicting cognitive decline in AD or MS patients. As a result, further studies using similar approaches on different cohorts will be required to enhance the generalization of these findings.

AUTHOR CONTRIBUTIONS

Borja Camino-Pontes: Formal analysis; investigation; methodology; visualization; writing – original draft; writing – review and editing. **Francisco Gonzalez-Lopez:** Data curation; formal analysis; investigation; methodology; visualization; writing – original draft. **Gonzalo Santamaría-Gomez:** Data curation; formal analysis. **Antonio Javier Sutil-Jimenez:** Investigation; methodology; writing – original draft; writing – review and editing. **Carolina Sastre-Barrios:** Resources; validation; writing – original draft. **Iñigo Fernandez de Pierola:** Project administration; resources; writing – original draft. **Jesus M Cortes:** Conceptualization; formal analysis; investigation; methodology; supervision; writing – original draft; writing – review and editing.

ACKNOWLEDGEMENTS

This work was funded by the Centro para el Desarrollo Tecnológico Industrial de España (CDTI) (grant no. EXP 00117191 / IDI-20190726), and Agencia de Desarrollo Económico de La Rioja ADER (2019-I-CHE-00020).

CONFLICT OF INTEREST STATEMENT

The precise formula of the NeuronUP Score, a different one for each TM and from which we defined class 1 (cognitive decline) versus class 0 (no cognitive decline) is protected by NeuronUP. As acknowledged in the Data Availability Statement, if favourable, the final s_0 and s_1 scores on which the analyses were performed will be available. Borja Camino-Pontes: No conflict of interest. Gonzalo Santamaría Gomez: Paid by NeuronUP with a contract associated to the Grant funded by Centro para el Desarrollo Tecnológico Industrial de España (CDTI) (grant no. EXP 00117191 / IDI-20190726) and Agencia de Desarrollo Económico de La Rioja ADER (2019-I-CHE-00020). Francisco Gonzalez-Lopez: Paid by NeuronUP with a contract associated to the Grant funded y Centro para el Desarrollo Tecnológico Industrial de España (CDTI) (grant no. EXP 00117191 / IDI-20190726). Antonio Javier Sutil-Jimenez: No conflict of interest. Carolina Sastre-Barrios: Neuropsychologist in NeuronUP Team. Iñigo Fernandez de Pierola: Neuropsychologist in NeuronUP Team; CEO. Jesus M Cortes: Consultancy Researcher and Responsible for the training of Francisco Gonzalez-Lopez and Gonzalo Santamaría Gomez within the grant funded by Centro para el Desarrollo Tecnológico Industrial de España (CDTI) (grant no. EXP 00117191 / IDI-20190726), Agencia de Desarrollo Económico de La Rioja ADER (2019-I-CHE-00020), and research and development funds allocated by NeuronUP from its own resources.

DATA AVAILABILITY STATEMENT

The authors agree to make available upon reasonable request fully anonymized data of the subjects' baseline score (s0) and 1-year after score (s1) for each instance and TM, in addition to percentiles of trues, time and level of difficulty of all different instances, needed for replication of all analyses presented in this article.

ORCID

Borja Camino-Pontes https://orcid.org/0000-0002-9071-9304
Jesus M. Cortes https://orcid.org/0000-0002-9059-8194

REFERENCES

Albert, M. S., Moss, M. B., Tanzi, R., & Jones, K. (2001). Preclinical prediction of AD using neuropsychological tests. *Journal of the International Neuropsychological Society*, *7*(5), 631–639. https://doi.org/10.1017/s1355617701755105

Arroyo-Ferrer, A., Sánchez-Cuesta, F. J., González-Zamorano, Y., del Castillo, M., Sastre-Barrios, C., Ríos-Lago, M., & Romero, J. P. (2021). Validation of cognitive rehabilitation as a balance rehabilitation strategy in patients with Parkinson's disease: Study protocol for a randomized controlled trial. *Medicina (Kaunas, Lithuania)*, *57*(4), 314. https://doi.org/10.3390/medicina57040314

Baiano, C., Barone, P., Trojano, L., & Santangelo, G. (2020). Prevalence and clinical aspects of mild cognitive impairment in Parkinson's disease: A meta-analysis. *Movement Disorders*, *35*(1), 45–54. https://doi.org/10.1002/mds.27902

Ben Miled, Z., Haas, K., Black, C. M., Khandker, R. K., Chandrasekaran, V., Lipton, R., & Boustani, M. A. (2020). Predicting dementia with routine care EMR data. *Artificial Intelligence in Medicine*, *102*, 101771. https://doi.org/10.1016/j.artmed.2019.101771

Bucholc, M., Ding, X., Wang, H., Glass, D. H., Wang, H., Prasad, G., Maguire, L. P., Bjourson, A. J., McClean, P. L., Todd, S., Finn, D. P., & Wong-Lin, K. F. (2019). A practical computerized decision support system for predicting the severity of Alzheimer's disease of an individual. *Expert Systems with Applications*, *130*, 157–171. https://doi.org/10.1016/j.eswa.2019.04.022

Casanova, R., Saldana, S., Lutz, M. W., Plassman, B. L., Kuchibhatla, M., & Hayden, K. M. (2020). Investigating predictors of cognitive decline using machine learning. *The Journals of Gerontology. Series B, Psychological Sciences and Social Sciences*, *75*(4), 733–742. https://doi.org/10.1093/geronb/gby054

Cohen, J. A., Trojano, M., Mowry, E. M., Uitdehaag, B. M., Reingold, S. C., & Marrie, R. A. (2020). Leveraging real-world data to investigate multiple sclerosis disease behavior, prognosis, and treatment. *Multiple Sclerosis Journal*, *26*(1), 23–37. https://doi.org/10.1177/1352458519892555

Davis, M., Connell, T. O., Johnson, S., Cline, S., Merikle, E., Martenyi, F., & Simpson, K. (2018). Estimating Alzheimer's disease progression rates from normal cognition through mild cognitive impairment and stages of dementia. *Current Alzheimer Research*, *15*(8), 777–788.

Dubois, B., Epelbaum, S., Nyasse, F., Bakardjian, H., Gagliardi, G., Uspenskaya, O., Houot, M., Lista, S., Cacciamani, F., Potier, M. C., Bertrand, A., Lamari, F., Benali, H., Mangin, J. F., Colliot, O., Genthon, R., Habert, M. O., Hampel, H., & INSIGHT-preAD Study Group. (2018). Cognitive and neuroimaging features and brain β-amyloidosis in individuals at risk of Alzheimer's disease (INSIGHT-preAD): A longitudinal observational study. *Lancet Neurology*, *17*(4), 335–346. https://doi.org/10.1016/S1474-4422(18)30029-2

Ewers, M., Walsh, C., Trojanowski, J. Q., Shaw, L. M., Petersen, R. C., Jack, C. R., Jr., Feldman, H. H., Bokde, A. L., Alexander, G. E., Scheltens, P., Vellas, B., Dubois, B., Weiner, M., Hampel, H., & North American Alzheimer's Disease Neuroimaging Initiative (ADNI). (2012). Prediction of conversion from mild cognitive impairment to Alzheimer's disease dementia based upon biomarkers and neuropsychological test performance. *Neurobiology of Aging*, *33*(7), 1203–1214.e2. https://doi.org/10.1016/j.neurobiolaging.2010.10.019

Ezzati, A., Zammit, A. R., Harvey, D. J., Habeck, C., Hall, C. B., Lipton, R. B., & Alzheimer's Disease Neuroimaging Initiative. (2019). Optimizing machine learning methods to improve predictive models of Alzheimer's disease. *Journal of Alzheimer's Disease*, *71*(3), 1027–1036. https://doi.org/10.3233/JAD-190262

Fields, J. A., Ferman, T. J., Boeve, B. F., & Smith, G. E. (2011). Neuropsychological assessment of patients with dementing illness. *Nature Reviews. Neurology*, *7*(12), 687. https://doi.org/10.1038/nrneurol.2011.173

Fouladvand, S., Mielke, M. M., Vassilaki, M., St, J., Sauver, R. C., Petersen, R. C., & Sohn, S. (2019). Deep learning prediction of mild cognitive impairment using electronic health records. *Proceeding of IEEE International Conference on Bioinformatics and Biomedicine*, *2019*, 799–806. https://doi.org/10.1109/bibm47256.2019.8982955

García-Herranz, S., Díaz-Mardomingo, M. C., & Peraita, H. (2016). Neuropsychological predictors of conversion to probable Alzheimer disease in elderly with mild cognitive impairment. *Journal of Neuropsychology*, *10*(2), 239–255. https://doi.org/10.1111/jnp.12067

Gavett, B. E., Ozonoff, A., Doktor, V., Palmisano, J., Nair, A. K., Green, R. C., Jefferson, A. L., & Stern, R. A. (2010). Predicting cognitive decline and conversion to Alzheimer's disease in older adults using the NAB list learning test. *Journal of the International Neuropsychological Society*, *16*(4), 651–660. https://doi.org/10.1017/S1355617710000421

Gleason, C. E., Norton, D., Anderson, E. D., Wahoske, M., Washington, D. T., Umucu, E., Koscik, R. L., Dowling, N. M., Johnson, S. C., Carlsson, C. M., Asthana, S., & Alzheimer's Disease Neuroimaging Initiative. (2018). Cognitive variability predicts incident Alzheimer's disease and mild cognitive impairment comparable to a cerebrospinal fluid biomarker. *Journal of Alzheimer's Disease*, *61*(1), 79–89. https://doi.org/10.3233/JAD-170498

Gomar, J. J., Conejero-Goldberg, C., Davies, P., & Goldberg, T. E. (2014). Extension and refinement of the predictive value of different classes of markers in ADNI: Four-year follow-up data. *Alzheimer's & Dementia*, *10*(6), 704–712. https://doi.org/10.1016/j.jalz.2013.11.009

Graham, S. A., Lee, E. E., Jeste, D. V., van Patten, R., Twamley, E. W., Nebeker, C., Yamada, Y., Kim, H. C., & Depp, C. A. (2020). Artificial intelligence approaches to predicting and detecting cognitive decline in older adults: A conceptual review. *Psychiatry Research*, *284*, 112732. https://doi.org/10.1016/j.psychres.2019.112732

Grande, G., Vanacore, N., Vetrano, D. L., Cova, I., Rizzuto, D., Mayer, F., Maggiore, L., Ghiretti, R., Cucumo, V., Mariani, C., Cappa, S. F., & Pomati, S. (2018). Free and cued selective reminding test predicts progression to Alzheimer's disease in people with mild cognitive impairment. *Neurological Sciences*, *39*(11), 1867–1875. https://doi.org/10.1007/s10072-018-3507-y

Irazoki, E., Contreras-Somoza, L. M., Toribio-Guzmán, J. M., Jenaro-Río, C., van der Roest, H., & Franco-Martín, M. A. (2020). Technologies for cognitive training and cognitive rehabilitation for people with mild cognitive impairment and dementia. A systematic review. *Frontiers in Psychology*, *11*, 648. https://doi.org/10.3389/fpsyg.2020.00648

Jaakkimainen, R. L., Bronskill, S. E., Tierney, M. C., Herrmann, N., Green, D., Young, J., Ivers, N., Butt, D., Widdifield, J., & Tu, K. (2016). Identification of physician-diagnosed Alzheimer's disease and related dementias in population-based administrative data: A validation study using family Physicians' electronic medical records. *Journal of Alzheimer's Disease*, *54*(1), 337–349. https://doi.org/10.3233/JAD-160105

Jarow, J. P., LaVange, L., & Woodcock, J. (2017). Multidimensional evidence generation and FDA regulatory decision making: Defining and using "real-world" data. *JAMA*, *318*(8), 703–704. https://doi.org/10.1001/jama.2017.9991

Kim, H., Chun, H.-W., Kim, S., Coh, B.-Y., Kwon, O.-J., & Moon, Y.-H. (2017). Longitudinal study-based dementia prediction for public health. *International Journal of Environmental Research and Public Health*, *14*(9), 983. https://doi.org/10.3390/ijerph14090983

Lee, D. Y., Youn, J. C., Choo, I. H., Kim, K. W., Jhoo, J. H., Pak, Y. S., Suh, K. W., & Woo, J. I. (2006). Combination of clinical and neuropsychologic information as a better predictor of the progression to Alzheimer disease in questionable dementia individuals. *The American Journal of Geriatric Psychiatry*, *14*(2), 130–138. https://doi.org/10.1097/01.JGP.0000192487.58426.d2

Lee, G., Nho, K., Kang, B., Sohn, K.-A., & Kim, D. (2019). Predicting Alzheimer's disease progression using multi-modal deep learning approach. *Scientific Reports*, *9*(1), 1952. https://doi.org/10.1038/s41598-018-37769-z

Lega, C., Cattaneo, L., & Costantini, G. (2022). How to test the association between baseline performance level and the modulatory effects of non-invasive brain stimulation techniques. *Frontiers in Human Neuroscience*, *16*, 920558. https://doi.org/10.3389/fnhum.2022.920558

Li, H., & Fan, Y. (2019). Early prediction of Alzheimer's disease dementia based on baseline hippocampal MRI and 1-year follow-up cognitive measures using deep recurrent neural networks. *Proceedings of IEEE International Symposium on Biomedical Imaging, 2019*, 368–371. https://doi.org/10.1109/ISBI.2019.8759397

Li, S., Okonkwo, O., Albert, M., & Wang, M.-C. (2013). Variation in variables that predict Progression from MCI to AD dementia over duration of follow-up. *American Journal of Alzheimer's Disease*, *2*(1), 12–28. https://doi.org/10.7726/ajad.2013.1002

Luo, H., Lau, K. K., Wong, G. H. Y., Chan, W. C., Mak, H. K. F., Zhang, Q., Knapp, M., & Wong, I. C. K. (2020). Predicting dementia diagnosis from cognitive footprints in electronic health records: A case-control study protocol. *BMJ Open*, *10*(11), e043487. https://doi.org/10.1136/bmjopen-2020-043487

Matthews, P. M., Block, V. J., & Leocani, L. (2020). E-health and multiple sclerosis. *Current Opinion in Neurology*, *33*(3), 271–276. https://doi.org/10.1097/WCO.0000000000000823

Mawdsley, E., Reynolds, B., & Cullen, B. (2021). A systematic review of the effectiveness of machine learning for predicting psychosocial outcomes in acquired brain injury: Which algorithms are used and why? *Journal of Neuropsychology*, *15*(3), 319–339. https://doi.org/10.1111/jnp.12244

Mendoza Laiz, N., Del Valle Díaz, S., Rioja Collado, N., Gomez-Pilar, J., & Hornero, R. (2018). Potential benefits of a cognitive training program in mild cognitive impairment (MCI). *Restorative Neurology and Neuroscience*, *36*(2), 207–213. https://doi.org/10.3233/RNN-170754

Milne-Ives, M., van Velthoven, M. H., & Meinert, E. (2020). Mobile apps for real-world evidence in health care. *Journal of the American Medical Informatics Association*, *27*(6), 976–980. https://doi.org/10.1093/jamia/ocaa036

Monti, S., Grosso, V., Todoerti, M., & Caporali, R. (2018). Randomized controlled trials and real- world data: Differences and similarities to untangle literature data. *Rheumatology*, *57*(Supplement_7), vii54–vii58. https://doi.org/10.1093/rheumatology/key109

Murueta-Goyena, A., del Pino, R., Galdós, M., Arana, B., Acera, M., Carmona-Abellán, M., Fernández-Valle, T., Tijero, B., Lucas-Jiménez, O., Ojeda, N., Ibarretxe-Bilbao, N., Peña, J., Cortes, J., Ayala, U., Barrenechea, M., Gómez-Esteban, J. C., & Gabilondo, I. (2021). Retinal thickness predicts the risk of cognitive decline in Parkinson disease. *Annals of Neurology*, *89*(1), 165–176. https://doi.org/10.1002/ana.25944

Murueta-Goyena, A., Pino, R., Reyero, P., Galdós, M., Arana, B., Lucas-Jiménez, O., Acera, M., Tijero, B., Ibarretxe-Bilbao, N., Ojeda, N., Peña, J., Cortés, J., Gómez-Esteban, J. C., & Gabilondo, I. (2019). Parafoveal thinning of inner retina is associated with visual dysfunction in Lewy body diseases. *Movement Disorders*, *34*(9), 1315–1324. https://doi.org/10.1002/mds.27728

Næss, K.-A. B., Lyster, S.-A. H., Hulme, C., & Melby-Lervåg, M. (2011). Language and verbal short-term memory skills in children with down syndrome: A meta-analytic review. *Research in Developmental Disabilities*, *32*(6), 2225–2234. https://doi.org/10.1016/j.ridd.2011.05.014

Nori, V. S., Hane, C. A., Martin, D. C., Kravetz, A. D., & Sanghavi, D. M. (2019). Identifying incident dementia by applying machine learning to a very large administrative claims dataset. *PLoS One*, *14*(7), e0203246. https://doi.org/10.1371/journal.pone.0203246

Oldham, P. D. (1962). A note on the analysis of repeated measurements of the same subjects. *Journal of Chronic Diseases*, *15*(10), 969–977. https://doi.org/10.1016/0021-9681(62)90116-9

Palmqvist, S., Hertze, J., Minthon, L., Wattmo, C., Zetterberg, H., Blennow, K., Londos, E., & Hansson, O. (2012). Comparison of brief cognitive tests and CSF biomarkers in predicting Alzheimer's disease in mild cognitive impairment: Six-year follow-up study. *PLoS One*, *7*(6), e38639. https://doi.org/10.1371/journal.pone.0038639

Parikh, M., Hynan, L. S., Weiner, M. F., Lacritz, L., Ringe, W., & Cullum, C. M. (2014). Simple neuropsychological test scores associated with rate of cognitive decline in early Alzheimer disease. *The Clinical Neuropsychologist*, *28*(6), 926–940. https://doi.org/10.1080/13854046.2014.944937

Pedersen, K. F., Larsen, J. P., Tysnes, O.-B., & Alves, G. (2013). Prognosis of mild cognitive impairment in early Parkinson disease: The Norwegian ParkWest study. *JAMA Neurology*, *70*(5), 580–586. https://doi.org/10.1001/jamaneurol.2013.2110

Pereira, T., Ferreira, F. L., Cardoso, S., Silva, D., de Mendonça, A., Guerreiro, M., Madeira, S. C., & Alzheimer's Disease Neuroimaging Initiative. (2018). Neuropsychological predictors of conversion from mild cognitive impairment to Alzheimer's disease: A feature selection ensemble combining stability and predictability. *BMC Medical Informatics and Decision Making*, *18*(1), 137. https://doi.org/10.1186/s12911-018-0710-y

Pertíñez, G. G., & Linares, A. G. (2015). Platforms for neuropsychological rehabilitation: Current status and lines of work. *Neurología (English Edition)*, *30*(6), 359–366. https://doi.org/10.1016/j.nrleng.2013.06.022

Ponjoan, A., Garre-Olmo, J., Blanch, J., Fages, E., Alves-Cabratosa, L., Martí-Lluch, R., Comas-Cufí, M., Parramon, D., García-Gil, M., & Ramos, R. (2019). How well can electronic health records from primary care identify Alzheimer's disease cases? *Clinical Epidemiology*, *11*, 509–518. https://doi.org/10.2147/CLEP.S206770

Ponjoan, A., et al. (2020). Is it time to use real-world data from primary care in Alzheimer's disease? *Alzheimer's Research & Therapy*, *12*(1), 60. https://doi.org/10.1186/s13195-020-00625-2

Ritchie, C. W., Khandker, R. K., Pike, J., Black, C. M., Jones, E., & Ambegaonkar, B. M. (2018). Real-world, multinational, retrospective observational survey of the ADAS-cog and associations with healthcare resource utilization in patients with Alzheimer's disease. *Journal of Alzheimer's Disease*, *64*(3), 899–910. https://doi.org/10.3233/JAD-180306

Sala, I., Illán-Gala, I., Alcolea, D., Sánchez-Saudinós, M. B., Salgado, S. A., Morenas-Rodríguez, E., Subirana, A., Videla, L., Clarimón, J., Carmona-Iragui, M., Ribosa-Nogué, R., Blesa, R., Fortea, J., & Lleó, A. (2017). Diagnostic and prognostic value of the combination of two measures of verbal memory in mild cognitive impairment due to Alzheimer's disease. *Journal of Alzheimer's Disease*, *58*(3), 909–918. https://doi.org/10.3233/JAD-170073

Seccia, R., Romano, S., Salvetti, M., Crisanti, A., Palagi, L., & Grassi, F. (2021). Machine learning use for prognostic purposes in multiple sclerosis. *Life*, *11*(2), 2. https://doi.org/10.3390/life11020122

Shehzad, A., Rockwood, K., Stanley, J., Dunn, T., & Howlett, S. E. (2020). Use of patient-reported symptoms from an online symptom tracking tool for dementia severity staging: Development and validation of a machine learning approach. *Journal of Medical Internet Research*, *22*(11), e20840. https://doi.org/10.2196/20840

Sherman, R. E., Anderson, S. A., Dal Pan, G. J., Gray, G. W., Gross, T., Hunter, N. L., LaVange, L., Marinac-Dabic, D., Marks, P. W., Robb, M. A., Shuren, J., Temple, R., Woodcock, J., Yue, L. Q., & Califf, R. M. (2016). Real-world evidence — What is it and what can it tell us? *The New England Journal of Medicine*, *375*(23), 2293–2297. https://doi.org/10.1056/NEJMsb1609216

Silverman, W. (2007). Down syndrome: Cognitive phenotype. *Mental Retardation and Developmental Disabilities Research Reviews*, *13*(3), 228–236. https://doi.org/10.1002/mrdd.20156

Sindi, S., Calov, E., Fokkens, J., Ngandu, T., Soininen, H., Tuomilehto, J., & Kivipelto, M. (2015). The CAIDE dementia risk score app: The development of an evidence- based mobile application to predict the risk of dementia. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, *1*(3), 328–333. https://doi.org/10.1016/j.dadm.2015.06.005

Solomon, A., & Soininen, H. (2015). Dementia: Risk prediction models in dementia prevention. *Nature Reviews. Neurology*, *11*(7), 375–377. https://doi.org/10.1038/nrneurol.2015.81

Tabarestani, S., Aghili, M., Eslami, M., Cabrerizo, M., Barreto, A., Rishe, N., Curiel, R. E., Loewenstein, D., Duara, R., & Adjouadi, M. (2020). A distributed multitask multimodal approach for the prediction of Alzheimer's disease in a longitudinal study. *NeuroImage*, *206*, 116317. https://doi.org/10.1016/j.neuroimage.2019.116317

Tatsuoka, C., Tseng, H., Jaeger, J., Varadi, F., Smith, M. A., Yamada, T., Smyth, K. A., Lerner, A. J., & The Alzheimer's Disease Neuroimaging Initiative. (2013). Modeling the heterogeneity in risk of progression to Alzheimer's disease across cognitive profiles in mild cognitive impairment. *Alzheimer's Research & Therapy*, *5*(2), 14. https://doi.org/10.1186/alzrt168

U.S. Food & Drug Administration. (2022). Real-World Evidence, *FDA*, 20 de mayo de. https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence

Vaccaro, M. G., Sarica, A., Quattrone, A., Chiriaco, C., Salsone, M., Morelli, M., & Quattrone, A. (2021). Neuropsychological assessment could distinguish among different clinical phenotypes of progressive supranuclear palsy: A machine learning approach. *Journal of Neuropsychology*, *15*(3), 301–318. https://doi.org/10.1111/jnp.12232

Wolfsgruber, S., Jessen, F., Wiese, B., Stein, J., Bickel, H., Mösch, E., Weyerer, S., Werle, J., Pentzek, M., Fuchs, A., Köhler, M., Bachmann, C., Riedel-Heller, S. G., Scherer, M., Maier, W., Wagner, M., & AgeCoDe Study Group. (2014). The CERAD neuropsychological assessment battery total score detects and predicts Alzheimer disease dementia with high diagnostic accuracy. *The American Journal of Geriatric Psychiatry*, *22*(10), 1017–1028. https://doi.org/10.1016/j.jagp.2012.08.021

Young, A. L., Oxtoby, N. P., Daga, P., Cash, D. M., Fox, N. C., Ourselin, S., Schott, J. M., Alexander, D. C., & Alzheimer's Disease Neuroimaging Initiative. (2014). A data-driven model of biomarker changes in sporadic Alzheimer's disease. *Brain*, *137*(Pt 9), 2564–2577. https://doi.org/10.1093/brain/awu176

Yuan, H., Ali, M. S., Brouwer, E. S., Girman, C. J., Guo, J. J., Lund, J. L., Patorno, E., Slaughter, J. L., Wen, X., Bennett, D., & The ISPE Comparative Effectiveness Research Special Interest Group. (2018). Real-world evidence: What it is and what it can tell us according to the International Society for Pharmacoepidemiology (ISPE) comparative effectiveness research (CER) special interest group (SIG). *Clinical Pharmacology and Therapeutics*, *104*(2), 239–241. https://doi.org/10.1002/cpt.1086

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Camino-Pontes, B., Gonzalez-Lopez, F., Santamaría-Gomez, G., Sutil-Jimenez, A. J., Sastre-Barrios, C., de Pierola, I. F., & Cortes, J. M. (2023). One-year prediction of cognitive decline following cognitive-stimulation from real-world data. *Journal of Neuropsychology*, *00*, 1–17. https://doi.org/10.1111/jnp.12307