Machine Learning Analysis of Digital Clock Drawing Test Performance for Differential Classification of Mild Cognitive Impairment Subtypes Versus Alzheimer's Disease

Article *in* Journal of the International Neuropsychological Society · March 2020

DOI: 10.1017/S1355617720000144

CITATIONS
0

READS
353

9 authors, including:

Nicholas Calzaretto
Stevens Institute of Technology
1 PUBLICATION **0** CITATIONS
SEE PROFILE

Jacob Ryan Epifano
Rowan University
1 PUBLICATION **0** CITATIONS
SEE PROFILE

Sean McGuire
Rowan University
2 PUBLICATIONS **1** CITATION
SEE PROFILE

Muhammad Umer
Rowan University
10 PUBLICATIONS **36** CITATIONS
SEE PROFILE

Journal of the International Neuropsychological Society (2020), 1–11

doi:10.1017/S1355617720000144

Machine Learning Analysis of Digital Clock Drawing Test Performance for Differential Classification of Mild Cognitive Impairment Subtypes Versus Alzheimer's Disease

Russell Binaco[1,†], Nicholas Calzaretto[1,†], Jacob Epifano[1,†], Sean McGuire[1,†], Muhammad Umer[1], Sheina Emrani[2], Victor Wasserman[2], David J. Libon[2,3] and Robi Polikar[1,*]

[1]Signal Processing and Pattern Recognition Laboratory, Rowan University Glassboro, Glassboro, NJ, USA
[2]Department of Psychology, Rowan University, Glassboro, NJ, USA
[3]New Jersey Institute for Successful Aging, School of Osteopathic Medicine, Rowan University, Stratford, NJ, USA

(Received July 16, 2019; Final revision December 17, 2019; Accepted January 1, 2020)

Abstract

Objective: To determine how well machine learning algorithms can classify mild cognitive impairment (MCI) subtypes and Alzheimer's disease (AD) using features obtained from the digital Clock Drawing Test (dCDT). **Methods:** dCDT protocols were administered to 163 patients diagnosed with AD($n = 59$), amnestic MCI (aMCI; $n = 26$), combined mixed/dysexecutive MCI (mixed/dys MCI; $n = 43$), and patients without MCI (non-MCI; $n = 35$) using standard clock drawing command and copy procedures, that is, draw the face of the clock, put in all of the numbers, and set the hands for "10 after 11." A digital pen and custom software recorded patient's drawings. Three hundred and fifty features were evaluated for maximum information/minimum redundancy. The best subset of features was used to train classification models to determine diagnostic accuracy. **Results:** Neural network employing information theoretic feature selection approaches achieved the best 2-group classification results with 10-fold cross validation accuracies at or above 83%, that is, AD *versus* non-MCI = 91.42%; AD *versus* aMCI = 91.49%; AD *versus* mixed/dys MCI = 84.05%; aMCI *versus* mixed/dys MCI = 84.11%; aMCI *versus* non-MCI = 83.44%; and mixed/dys MCI *versus* non-MCI = 85.42%. A follow-up two-group non-MCI *versus* all MCI patients analysis yielded comparable results (83.69%). Two-group classification analyses were achieved with 25–125 dCDT features depending on group classification. Three- and four-group analyses yielded lower but still promising levels of classification accuracy. **Conclusion:** Early identification of emergent neurodegenerative illness is criterial for better disease management. Applying machine learning to standard neuropsychological tests promises to be an effective first line screening method for classification of non-MCI and MCI subtypes. (*JINS*, 2020, *xx*, xx-xx)

Keywords: Clock drawing, The Digital Clock Drawing Test, Mild cognitive impairment, Machine learning, Cognitive assessment

INTRODUCTION

Alzheimer's disease (AD) is a major public health problem; by 2050, it is estimated that 14 million people may be affected by this illness (Alzheimer's Association, 2019). There is considerable interest in the diagnosis and characterization of mild cognitive impairment (MCI), a syndrome believed to be a prodrome often leading to dementia. A variety of MCI subtypes have been identified, including patients presenting with single-domain syndromes such as amnestic (aMCI) and dysexecutive (dys MCI) and a combined or mixed (mixed MCI) phenotype (Emrani et al., 2019; Eppig et al., 2012; Libon et al., 2011). Past research suggests varying rates of conversion to dementia and reversion to normal cognitive functioning depending on MCI subtype (Aerts et al., 2017; Díaz-Mardomingo, García-Herranz, Rodríguez-Fernández, Venero, & Peraita, 2017; Pandya, Lacritz, Weiner, Deschner, & Woon, 2017; Shimada, Doi, Lee, & Makizako, 2019). There is some suggestion that MCI subtypes are associated with different underlying biological mechanisms. Thus, as disease-modifying agents become available, the differentiation between MCI subtypes will become increasingly important for disease management,

*Correspondence and reprint requests to: Robi Polikar, PhD, Signal Processing and Pattern Recognition Laboratory, Rowan University, Glassboro, NJ, USA. E-mail: polikar@rowan.edu
†These authors contributed equally to the work.

better understanding of the course of the illness, and predicting possible treatment outcomes. The development of machine learning-assisted tools for detecting emerging neurodegenerative illness with appropriate specificity may be able to address these needs.

The Clock Drawing Test (CDT) is a popular and widely used neuropsychological test (Rabin, Barr, & Burton, 2005). Over the past several decades, there has been considerable research regarding differing patterns of performance on the CDT among patients with various dementia and MCI syndromes (Ahmed et al., 2016; Cahn-Weiner et al., 2003; Cosentino, Jefferson, Chute, Kaplan, & Libon, 2004; Freedman et al., 1994; Libon, Malamut, Swenson, Sands, & Cloud, 1996; Price et al., 2011; Rouleau, Salmon, Butters, Kennedy, & McGuire, 1992; Royall, Cordes, & Polk, 1998; Royall, Palmer, & Markides, 2017; Schillerstrom et al., 2007; Shulman, 1993; Tuokko, Hadjistavropoulos, Miller, & Beattie, 1992). In the aggregate, prior research has demonstrated that clock drawing to command versus copying a model of a clock produces complementary, but different patterns of performance associated with different underlying neurocognitive constructs (Ahmed et al., 2016; Cosentino et al., 2004; Libon et al., 1996; Price et al., 2011).

The CDT is also used as part of other tests such as the Montreal Cognitive Assessment (Nasreddine et al., 2005) and the Mini-Cog (Holsinger et al., 2015), tests that are now commonly used to screen for neuropsychological impairment in primary health settings. Brief screening measures for cognitive change have previously been noted for their limited intra-individual reliability (Feeney et al., 2016). Additionally, cognitive screening measures have been shown to be influenced by baseline intelligence measured on word-reading tasks, indicating that an individual's performance on traditional screening measures should be considered with respect to their premorbid abilities (Dykiert et al., 2016). To the extent that the CDT is used to screen for cognitive impairments, additional potential limitations include the subjectivity of various scoring systems, and the ability to score or capture only a small or limited number of features that might reveal the presence of early or emergent cognitive impairment.

A digital version of the CDT has recently been developed at the Lahey Clinic and Massachusetts Institute of Technology (Davis et al., 2010; Penney, Davis, et al., 2010; Penney, Libon, et al., 2010) that combines traditional paper and pen administration procedures using a digital pen and software that captures the patient's drawing in real time, allowing the observation of previously difficult-to-capture behavior. This technology has been used with both patients and healthy research participants. Libon et al. (2014) administered the digital Clock Drawing Test (dCDT) to patients with multiple sclerosis (MS) to investigate the bradyphrenia often seen in this patient group. In both the command and copy conditions, MS patients produced slower latencies during the latter portion of their drawings, suggesting problems with sustaining attention and concentration. Cohen et al. (2014) obtained dCDT protocols from patients diagnosed with major depression and found that younger patients spent a smaller proportion of time actually drawing, relative to non-drawing time, compared to the older depressed group. These data were thought to reflect the presence of excessive rumination in patients with depression. Lamar et al. (2016) examined non-demented, non-depressed adults who were grouped on the basis of whether the four cardinal anchor digits (i.e., the numbers 12, 6, 3, 9) were initially drawn before other digits. Local-level connectome analyses found that anchorers demonstrated superior efficiency in the left medial orbitofrontal and transverse temporal cortices as well as the right rostral anterior cingulate and superior frontal gyrus versus non-anchorers. Hierarchical modularity analyses suggested a higher degree of modular integration involving heteromodal regions of the ventral visual processing stream for anchorers versus non-anchorers.

Using machine learning algorithms, Davis et al. (2014) demonstrated that the dCDT was able to differentiate patients with AD and other dementia syndromes from normal controls (NCs). Finally, Souillard-Mandar et al. (2016) extracted digital clock drawing data from healthy controls, patients with MCI, patients with several dementia syndromes, and other groups and demonstrated impressive classification rates using several machine learning techniques. In sum, past research using the dCDT has demonstrated that this technology is able to measure very subtle and discrete behavior involving latency, decision-making, and graphomotor output – behavior previously unobtainable – to aid in the differential classification between dementia subtypes.

A question that has not yet been explored is how well the dCDT can differentiate between patients classified with various MCI subtypes and dementia such as AD when features extracted from the test are used to train an automated machine learning-based classifier. The current research aims to address this issue. In the current research, the dCDT was administered to memory clinic patients meeting criteria for non-MCI, amnestic MCI, and a combined mixed/dys MCI using Jak, Bondi criteria (2009), and patients diagnosed clinically with AD. Our goal was to assess how well machine learning data analytic algorithms – trained with optimally selected features mathematically deemed to be most informative – are able to classify patients into their respective groups. The ability of machine learning algorithms to differentiate between patient groups could potentially increase the usefulness of the CDT to screen for neuropsychological impairment in primary care or other healthcare settings.

METHODS

Participants

The study cohort consisted of 163 patients (92 female; 162 Caucasian). All patients were evaluated at The Memory Assessment Program, Rowan University, New Jersey Institute for Successful Aging; and at the Department of Neurology, Drexel University, Pennsylvania. Rowan University and Drexel University institutional review boards approved this investigation and the current research complied with the Declaration of Helsinki. All data were collected between

Table 1. Demographic information (means and standard deviations)

	Non-MCI ($n = 35$)	aMCI ($n = 26$)	Mixed/dys MCI ($n = 43$)	AD ($n = 59$)
Age (*SD*)	77.10 (6.66)	75.85 (6.57)	75.48 (7.78)	79.71 (5.34)
Gender (%female)	51.43	76.92	55.81	50.85
Education (*SD*)	14.46 (2.85)	14.54 (2.75)	13.07 (3.00)	13.45 (2.91)
MMSE (*SD*)	27.24 (3.15)	26.58 (2.44)	26.77 (1.84)	22.96 (2.87)

Note: With the exception of one, all study participants were Caucasian.

2005 and 2018. The work-up for MCI and dementia was identical between research sites and included comprehensive neuropsychological assessment, evaluations provided by a board-certified geriatric psychiatrist or neurologist, a brain MRI/CT scan, and serum blood tests. An interdisciplinary team consisting of a neuropsychologist, psychiatrist/or neurologist, and social worker determined clinical diagnosis. Exclusion criteria included any lifetime history of head injury, substance abuse, major/medical psychiatric disorders (e.g., major depression and epilepsy), B12, folate, or thyroid deficiency. The presence of these issues was determined on the basis of information provided by both patients and their families and was deemed positive if any of these issues interfered with work or other aspects of daily life. Any changes in functional independence in connection with a blow or strike to the head were considered indicative of history of head injury. All neuropsychological evaluations were obtained by one of us (DJL, a PhD level, broad-certified neuropsychologist). A knowledgeable family member provided information regarding functional status. Demographic and clinical characteristics including age, education, and Mini-Mental State Exam Performance (MMSE; Folstein, Folstein, & McHugh, 1975) are displayed in Table 1.

Neuropsychological Assessment

The neuropsychological protocol used to clinically diagnose non-MCI and MCI patients is identical as described by Emrani et al., (2018). Three domains of neuropsychological functioning were assessed including executive control, naming/lexical access, and verbal episodic memory. Nine neuropsychological parameters, three from each neurocognitive domain, were employed to classify patients as presenting with non-MCI or MCI subtypes as described below. Tests were expressed as *z*-scores derived from normative data as listed below.

Executive control

The Boston Revision of the Wechsler Memory Scale-Mental Control subtest-Accuracy Index (see Lamar, Price, Cynthia, Kaplan, & Libon, 2002 for full details); letter verbal fluency ('FAS'; Spreen & Strauss, 1990); and Trail Making Test-Part B (Reitan & Wolfson, 1985).

Lexical access/language

The 60-item Boston Naming Test (Kaplan, Goodglass, & Weintraub, 1983); semantic ('animals') fluency (Carew, Lamar, Cloud, Grossman, & Libon, 1997); and Wechsler Adult Intelligence Scale-III Similarities subtest (Wechsler, 1997).

Memory and learning

Episodic memory was assessed with three parameters obtained from the 9-word California Verbal Learning Test short form (Delis, Kramer, Kaplan, & Ober, 2000) including total immediate free recall, delayed free recall, and the delayed recognition discriminability index.

Diagnostic Determination

Alzheimer's disease

All AD patients were assessed at Rowan University Memory Assessment Program and diagnosed using the criteria suggested by McKhann et al. (1984).

Single- and multi-domain MCI

Jak, Bondi et al. (2009) comprehensive neuropsychological criteria were used to determine MCI subtype. Single-domain MCI syndromes were diagnosed when scores fell below 1.0 standard deviation (*SD*) below expectations based on available norms on any two of the three measures within a single cognitive domain. Mixed MCI syndromes were diagnosed when scores fell below 1.0 *SD* below available norms on any two of the three measures across two or more cognitive domains. Twenty-six patients met criteria for single-domain aMCI, fifteen patients for dys MCI, and twenty-eight for mixed or multi-domain mild cognitive impairment (mixed MCI). Because of the small number of dys MCI patients, a combined mixed/dys MCI subgroup ($n = 43$) was constructed. Supporting this decision, previous research employing these criteria has shown that statistically defined mixed and single-domain dys MCI patients produce similar patterns of performance on executive tests (Eppig et al., 2012).

Non-MCI group

Thirty-five patients did not meet Jak, Bondi et al. (2009) criteria for MCI. Some of these individuals ($n = 11$) obtained scores where all nine neuropsychological parameters were above 1 *SD* below mean cutoff criteria. The second group of patients ($n = 24$) not meeting criteria for MCI presented with some, but very little, cognitive impairment, such that 14 patients produced tests scores in which only one of the nine neuropsychological parameters was below 1 *SD* and 10 patients produced neuropsychological test scores in which only two parameters across different domains of cognitive functioning were below 1 *SD*. These patients were combined into a single group and labeled as non-MCI.

The Digital Clock Drawing Test

The instructions used to administer the dCDT are consistent with traditional CDT administration and included both a command and copy test conditions. In the command condition, patients were asked, "draw the face of a clock, put in all numbers, and set the hands for 10 after 11." Upon completion of the command test condition, the copy test condition was administered whereby patients were asked to copy a model of a clock with hands set for 10 after 11. The dCDT (Davis et al., 2010; Penney, Davis, et al., 2010) was developed by the Lahey Clinic and Massachusetts Institute of Technology in collaboration with the ClockSketch Consortium and uses digital pen technology developed by Anoto, Inc., Westborough, MA, USA. The pen works as an ordinary ball point pen while capturing pen position 80 times/sec. + .002. All data are time-stamped, allowing the pen to capture the final drawing for more accurate classification than would be possible without this technology. The Clocksketch program can label and calculate latencies for the length and number of strokes associated with individual clock components such as digits and hands; the dimensions and orientation of the clock face, digits, and hands; the time elapsed during and between the drawing of individual clock features; and the deviation of clock features from "ideal" placement on the command clock stimulus.

Data Preprocessing

Prior to formal analysis, all dCDT data underwent preprocessing to remove duplicate features, noisy features, and extremely sparse features, that is, those that are only available for a very small subset of the patients (such as clock drawing center dot and related features). All feature values were standardized by subtracting the mean and dividing with the *SD* of that feature obtained from the total sample. Standardization allows the down-stream processing to be independent of scaling and amplitude.

Information theoretic approaches for feature selection evaluate mutual information (MI) among classifier variables to identify and reduce the number of features with shared predictive contributions. These approaches were based on discrete outcome probabilities, that is, to augment the selection of predictive features, continuous values were converted to discrete, categorical values in order to better isolate MI contributed by these features. All continuous valued features were discretized by binning data into 10 bins using the discretization function in Gavin Brown's FEAST toolbox for Matlab® (Brown, 2009). The process of binning divides continuous values into $N = 10$ categorical values, where all N bins have equal ranges.

For some patients, features of interest were missing such as the absence of a clock hand or digit within the clock face. To resolve issues revolving around missing features, a *k*-nearest neighbor (kNN) algorithm was used to impute the missing values. For any missing feature, kNN searches the *k nearest* neighbors of available features with respect to Euclidean distance and averages them to fill in for the missing data (Beretta & Santaniello, 2016). To find kNNs, Euclidean distance was calculated for all relevant features. For example, if a patient did not draw the number "3," a search was performed on all patients of the same diagnosis who did draw the "3." Then, the *k* patients who were most similar with respect to the Euclidean distance to the patient in question were identified. The features associated with the number "3" were averaged from these *k* patients and were then incorporated as the missing values. To account for patients who did not draw an entire digit and/or clock hand(s), binary variables were created and coded such that a one indicated the value was originally missing and a zero indicated it was available. These additional features preserved the information that the value was imputed for the corresponding patient.

The last preprocessing step involved data partitioning with respect to diagnostic labels. As stated above, the overall goal of this project was to classify non-MCI, MCI subtypes, and AD patients into their respective groups. A singe four-group classification and two combinations of three-group classifications were calculated. However, because of modest sample size among our four patient groups, we were particularly interested in how well machine learning feature selection was able to classify patients into any combination of two groups. Hence, six 2-class problems were evaluated (i.e., all pairwise combinations of two diagnostic groups): non-MCI *versus* aMCI; non-MCI *versus* mixed/dys MCI; non-MCI *versus* AD; aMCI *versus* mixed/dys MCI; aMCI *versus* AD; and mixed/dys MCI *versus* AD. In performing these experiments, information theoretic feature selection algorithms were applied to the problem and then fit to an appropriate feedforward neural network classifier (see Table 2). In an additional, follow-up analysis, aMCI and mixed/dys MCI groups were combined to enable an analysis of classification accuracy for non-MCI *versus* MCI.

Information Theory-Based Feature Selection

The classification of patients through machine learning-based analysis of the dCDT was based on two processes:

Table 2. Information theoretic final results

	FS criterion	Neural network size	Number of features selected	Performance (%)	95% CI
Non-MCI *versus* aMCI	CMIM	2 L, 10, 10 N	25	84.33	7.03
Non-MCI *versus* mxMCI	CMIM	1 L, 50 N	25	85.42	9.11
Non-MCI *versus* MCI (combined phenotype)	MRMR	1 L, 50 N	100	83.69	9.85
Non-MCI *versus* AD	CMIM	2 L, 20, 10 N	125	91.42	6.02
aMCI *versus* mxMCI	MRMR	2 L, 10, 10 N	75	84.11	5.90
aMCI *versus* AD	MRMR	2 L, 20, 10 N	100	91.49	5.99
mxMCI *versus* AD	JMI	2 L, 20, 10 N	125	84.05	6.14
Non-MCI *versus* aMCI *versus* mxMCI	JMI	1 L, 50 N	125	71.64	6.46
aMCI *versus* mxMCI *versus* AD	MI	1 L, 50 N	50	75.97	6.19
Non-MCI *versus* aMCI *versus* mxMCI *versus* AD	MI	1 L, 50 N	100	64.05	4.92

FS criterion = feature selection criterion; mxMCI = mixed/dysexecutive mild cognitive impairment.

information theory-based feature selection to identify the most relevant constellation of clock features for distinguishing between groups and successive training of a neural network classifier to identify group membership based on training data for which the group was known.

As noted above, 350 command and copy dCDT features were obtained from the entire sample of 163 patients. Information theoretic-based feature selection approaches were used to determine the most informative features (Brown, 2009). Information theoretic methods evaluate different forms of entropy-based MI between diagnoses and features.

Several information theoretic measures are available for computing the amount of information carried by a set of attributes. In this study, we evaluated MI, minimum redundancy maximum relevancy (MRMR), joint mutual information (JMI), and conditional mutual information maximization (CMIM), as described in more detail below (Brown, 2009). All of these metrics are based on the central measure of information, defined by entropy, which measures the amount of uncertainty in a random variable, that is, the higher the uncertainty, the higher the information. Given a random variable X (e.g., a specific feature), the entropy of X is given by Equation 1.

$$H(X) = -\sum_{x\in X} p(x)\log p(x) \tag{1}$$

where $p(x)$ is the probability that the random variable X assumes the specific value x. For example, if the specific value of a feature is always the same (i.e., there is no uncertainty), the $p(x)$ is either 0 or 1, depending on the value of x.

Shannon's mutual information I (X,Y)

This is defined as the dependence between two random variables X and Y as shown in Equation 2.

$$I(X;Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right) \tag{2}$$

where $p(x, y)$ is the joint probability that the random variable X assumes a particular value x, and the random variable Y assumes a particular value y. For a given feature represented by the random variable X and corresponding diagnosis represented with the random variable Y, the MI between X and Y is a measure of the change in entropy (drop in uncertainty) of Y due to the presence of X. Therefore, high MI between a feature and a diagnosis indicates that the feature is a good predictor of its respective diagnosis.

Minimum redundancy maximum relevancy

This was defined below in Equation 3 and includes the consideration of redundancy in addition to maximizing the MI of a feature and a diagnostic label.

$$J_{mrmr}(X_n) = \mathrm{I}(X_n;Y) - \frac{1}{n-1}\sum_{k=1}^{n-1} I(X_n;X_k) \tag{3}$$

Specifically, the MI between two features X_n and X_k (i.e., the redundancy in knowing both X_n and X_k), averaged over all features $k = 1, \ldots, n-1$, is subtracted from the MI between the feature of interest X_n and the label Y. The purpose of MRMR is to eliminate redundant features, while still selecting a set of features that enable an accurate prediction of the class label.

Joint mutual information

As shown in Equation 4, this measure includes an additional positive term, *the conditional redundancy*, compared to MRMR.

$$J_{jmi} = I(X_n; Y) - \frac{1}{n-1}\sum_{k=1}^{n-1}[I(X_n; X_k) - I(X_n; X_k|Y)] \quad (4)$$

The conditional redundancy is the redundancy between two features X_n and X_k given the diagnostic label Y. This term is an indication of first-order interaction between those two features that, when used together, is beneficial in the prediction of the diagnostic group.

Finally, *CMIM* is similar to JMI; however, CMIM takes a more pessimistic approach than JMI (Equation 5).

$$J_{cmim} = I(X_n; Y) - max_k[I(X_n; X_k) - I(X_n; X_k|Y)] \quad (5)$$

Instead of taking the average of all first-order interactions (i.e., difference between redundancy and conditional redundancy), CMIM only considers the specific pair $(X_n\ ;\ X_k)$ whose first-order interaction outputs the maximum score. For this reason, CMIM selects the feature X_k based on the interaction with an already selected feature X_n that gives the lowest score J. Therefore, if X_k has high redundancy and low conditional redundancy with only one other feature, then X_k will have a low score regardless of the interactions of X_k with other features.

Using these metrics as a feature selection approach involves computing the relevant information theoretic metric for each of the features followed by rank ordering the features in descending order according to the scores they receive. Then, a subset of these features with the highest scores is selected as the most informative feature subset.

Balancing the Data

In the current data set, the largest group (i.e., patients with AD) had 59 patients and the smallest group (aMCI) has 26 patients. The adverse impact of imbalanced datasets, creating a bias in the classification decisions of most machine learning algorithms, is well known. To address the imbalanced data concern, we used the well-established Synthetic Minority Oversampling Technique (SMOTE) to rebalance the classes (Chawla, Bowyer, Hall, & Kegelmeyer, 2002). This approach adds carefully constructed synthetic representative data points to underrepresented classes near the vicinity of existing instances. SMOTE does this by first determining the kNNs of each existing minority instance (k is set to five by default). The algorithm then randomly chooses an instance of a minority class and randomly selects one of its kNNs. Synthetic instances are then created through interpolation along the vector space between the randomly selected instance and its kNN. This process is repeated for additional instances of each minority class until all the minority classes are the same size as the majority class. It is important to note that SMOTE was only used to rebalance training datasets with which the classification models were trained. All diagnostic performances listed in the Results section were computed on a test/ evaluation dataset that strictly consisted of real patient data only.

Classifier Training

Group classification of patients based on dCDT feature selection was accomplished using feed-forward neural networks, due to their previously reported success in classifying medical data for diagnostic purposes (Amato et al., 2013) and because of the ability to modify the number of hidden layers and the number of nodes per hidden layer, which allows the network to adjust its learning capacity and approximate decision boundaries (Autio, Juhola, & Laurikkala, 2007). To address over-fitting, we implemented a relatively "small" network with fewer adjustable weights. A gradient descent-based stochastic optimization approach was the optimizer used, with early stopping as a further protection against over-fitting (Kingma & Ba, 2014).

Three different neural network architectures were compared to identify the best predictive configuration for accurately identifying group membership based on dCDT features alone: (1) one hidden layer of 50 nodes, (2) two hidden layers of 10 nodes each, and (3) two hidden layers with 20 and 10 nodes. Each of these structures was trained for features selected with each of the four information theoretic criteria in order to determine the highest performer of each test case (see Supplemental Tables 1–9). We further implemented a 10-fold cross-validation to obtain a better prediction of the true generalization performance of the diagnostic classifier, as well as a confidence interval of the estimated performance. In all cases, the diagnostic performance was computed on test datasets that had been set aside and not used for training, and strictly consisting of real patient data only.

There is, of course, empirical interest in describing the contributions of individual features to accurate classification. Unfortunately, feed-forward machine learning analysis does not yield this kind of information, because it does not analyze data using a General Linear Model (GLM)/regression-based model. While magnitude is considered, there is no directionality assigned to features or the score that ranks features using MI analysis. In addition, any weight that the analysis assigns to a given feature is masked by the hidden layers of the neural network architecture. It would be inaccurate to think of the relationship between a given feature and a classification label as linear, that is, possessing an easily interpreted directional value, because all features are being considered holistically within the model to arrive at a "best fit" classification. While the features themselves and their rank order allow for speculation about what value individual features contribute to the accuracy of classification, much of what machine learning analyses extract from a given feature is hidden within the analysis, thus limiting anything that can be said beyond that a given feature is relevant to distinguishing between two or more groups.

RESULTS

Groups did not differ for education ($F[3, 161] = 2.33, p < .076, \eta^2 = .079$). AD patients were older ($F[3, 161] = 4.20, p < .001, \eta^2 = .040$) compared only to the mixed MCI group

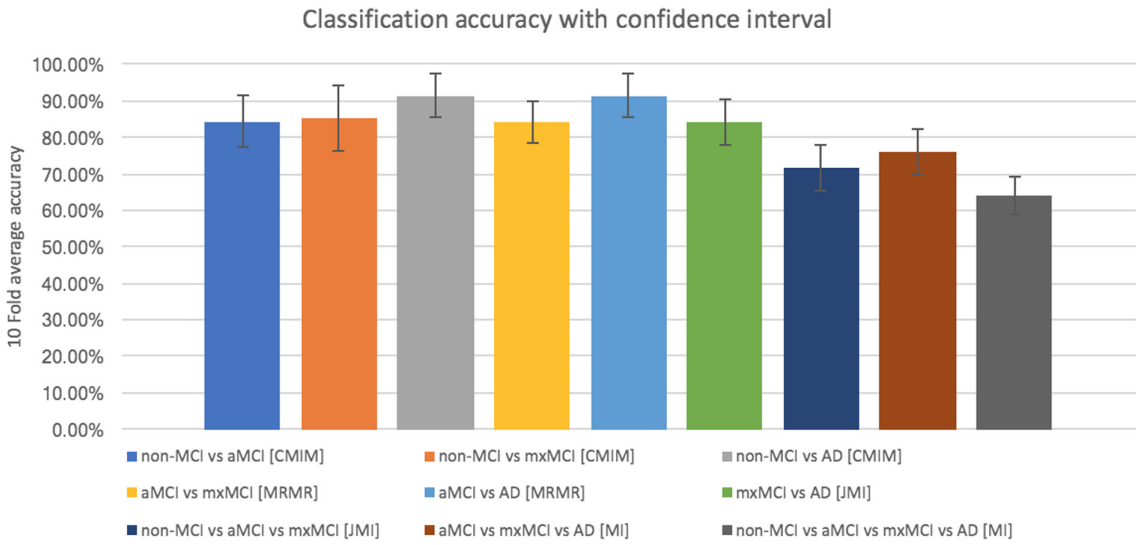

Fig. 1. Highest performing results for each test case using information theoretic feature selection criteria (see Table 1 for values). Error bars correspond to 95% CI.

($p < .009$). AD patients obtained lower scores on the MMSE ($F[3, 161] = 27.39$, $p < .001$, $\eta^2 = .345$) than all other groups (Table 1). There were no differences on the MMSE between the non-MCI and MCI groups. The results of all four-group, three-group, and two-group analyses are presented in Table 2 and Figure 1. The least complex model, that is, the model with the fewest adjustable weights or smallest network that obtained the best performance with a narrow confidence interval using as few features as possible, was chosen as the best result in each test case.

Figure 1 shows all 2-, 3-, and 4-diagnostic group analyses where the classifiers were trained with features selected based on information theoretic feature selection. As expected, each two-group analysis performed better as compared to three- and four-group comparisons. All two-group analyses surpassed 80% diagnostic accuracy in correctly classifying patients into their respective groups. Three- and four-group analyses yielded classification rates below 80%. Supplemental Tables 1–9 provide additional information, showing diagnostic performances obtained by each of the three different neural networks when trained with features obtained from each of the four information theoretic approaches discussed above. Supplemental Table 10 lists the first 20 features that entered into several two- and three-group analyses. A follow-up analysis was conducted whereby all MCI patients were placed in a single MCI group and compared to non-MCI patients. Feature section employed MRMR. Consistent with other two-group comparisons, this analysis also surpassed 80% diagnostic accuracy (83.69%; see Table 2).

DISCUSSION

The goal of the current research was to determine how well machine analytic techniques could classify AD patients, non-MCI patients, amnestic MCI, and combined mixed/dys MCI patents into their respective subtypes. Prior research has demonstrated that non-MCI and selected MCI subtypes can be differentiated using tests assessing mental manipulation (Emrani et al., 2018), verbal versus visual working memory (Emrani et al., 2019), verbal versus visual episodic memory (Wasserman et al., 2019), and selected visuospatial operations (Wasserman et al., submitted). These data are compelling on the basis of the neurocognitive constructs that are assessed and differentiated between groups. However, the complexity of the tests used in this research are not necessarily appropriate to be deployed as front-line screening measures.

No single neuropsychological test, including the dCDT, can substitute for detailed neuropsychological assessment in diagnosing dementia, MCI, or any other neurological condition. Moreover, research into how well the CDT can screen for dementia is mixed. Brodaty and Moore (1997) reported that the CDT was superior to the MMSE in screening for dementia, such as AD, in patients evaluated in a memory clinic. Other researchers assessing participants drawn from primary care found mixed results regarding how well the CDT was able to identify patients with dementia or MCI (Ehreke et al., 2009; Kirby et al., (2001). In a large population study comprised of over 6,000 participants drawn from primary care settings, the CDT was better than other neuropsychological screening measures in identifying participants with possible dementia (Reinera et al., 2018). Caballero et al. (2018) found that worse diabetes medication adherence was related to greater impairment on the CDT. Hetland et al. (2014) noted that older persons taking medication believed to impair driving skills produced greater impairment on the CDT.

Despite inconsistencies in past research, the results of the current research suggest that *digitally* obtained clock drawing behavior analyzed using machine learning algorithms may be able to screen for subtle to mild neuropsychological impairment. Subtle to mild neuropsychological impairment as seen

on the dCDT could, indeed, signal the emergence of a neurodegenerative illness such as AD. However, other aetiologies underlying subtle to mild neuropsychological impairment, such as psychiatric problems, might also result in impairment on the dCDT, including depression and anxiety, subtle medication side effects, and medical illness other than putative dementia. These questions should be the subject of future research. Nonetheless, deploying dCDT in the primary care environment and/ or other health care points of entry has potential for better overall healthcare delivery.

In the current research, we were able to demonstrate that feature selection coupled with a judiciously selected network architecture proved to be effective, particularly for two-way classification with performances in the 80%–90% range. For most of our results, the confidence intervals of different methods overlapped, indicating that different classifiers and feature selection approaches performed similarly. This is, perhaps, not surprising due to the similarity of each of the information theoretic metric, all of which are based on MI. It is worth repeating that, in general, the binary two-group analyses significantly outperformed the multi-class 3 and 4 group classification analyses. However, this outcome must be considered in the proper context of the difficulty of the problem, such that a classifier is said to provide meaningful utility by making intelligent decisions if its performance is better than random guessing. In a binary classification problem, that threshold is 50%; hence, classifiers performing in the 80%–90% range are clearly making intelligent decisions. In our four-group problem, however, that threshold is 25%. Thus, classifiers performing in the 60%–70% range are, in fact, providing meaningful information. While the three- and four-group problems performed more modestly, the results described above are nevertheless quite promising. However, we acknowledge that better multi-group classification rates are desirable. Additional research using a larger sample size may address these issues.

Classifying MCI subtypes and differentiating MCI from dementia such AD require a lengthy protocol including comprehensive neuropsychological tests, structural and/or biomarker-related CT/MRI scans of the brain, and an appropriate medical evaluation. As stated above, the current research does not suggest that the dCDT can replace the standard workup for MCI/ dementia nor are we suggesting the performance on a single test is sufficient to make an informed diagnosis. Nonetheless, the accuracies achieved for all binary problems are all at least 83%. As previously discussed, weaknesses in currently employed brief cognitive screening measures, including inconsistent test–retest reliability and sensitivity to premorbid intellect, leave such measures vulnerable to misclassifying patients. While the potential for misclassification can never be completely eliminated, the accuracy of classification achieved through machine learning analysis of a digitized, brief screening measure such as the dCDT may be a valuable indicator of emerging or undetected cognitive change. Thus, the administration of the dCDT is an inexpensive and efficient means to *screen* for possible alterations in neuropsychological abilities.

One of the strengths of the current study is the well-validated comprehensive neuropsychological diagnostic criteria used to classify non-MCI and MCI patients into their respective groups (Jak, Bondi et al., 2009). However, limitations must be acknowledged, including lack of ethnic diversity in our sample, possible bias in that AD patients were drawn from one site, and our modest sample size relative to the number of features that were analyzed. Also, although groups did not differ for education, premorbid abilities can influence performance on the CDT (Ainslie, et al., 1993). Future research exploring relations between premorbid abilities as related to group classification using machine-learning algorithms should be undertaken before a measure such as the dCDT can be used to screen for possible neuropsychological impairment. Additionally, the results of the current research were obtained from memory clinic patients. Different findings may have emerged when training a machine-learning classifier with a community-based sample. Future research should also recruit a community dwelling, NC group and assess how well machine-learning algorithms can classify NC versus any patient with MCI in addition to MCI subgroups.

Finally, the CDT is quite multidimensional regarding the neuropsychological abilities that are needed for successful performance. The methods used in the current research are not able to point to specific neuropsychological constructs that may underlie group classification. Future research combining machine learning classification followed by the analysis of algorithm-selected dCDT features and neuropsychological test performance may be able to provide information regarding the constructs that appear to contribute to group classification.

Additional future work also includes the need to use different feature selection approaches other than information theory, such as wrappers, genetic algorithms, and ensemble-based feature selections. However, it should be noted that each of these approaches are considerably more costly with respect to computational complexity and, in general, rely on specific classification algorithms. The information theoretic approaches, on the other hand, are independent of the classifier, providing a more unbiased estimate of the relevance of the features, not to mention their significantly lower computational cost. Perhaps a more interesting analysis is to compare a variety of feature selection approaches and then determine the overlap in the set of features selected among them. Such an analysis may provide additional insight into which features appear to be selected more often and are, perhaps, more informative than the others. Despite these limitations, machine-learning analysis of features selected from the dCDT is a promising tool to assess for the emergent neuropsychological impairment.

ACKNOWLEDGEMENTS

Dr. Libon owns stock in Digital Cognition Technologies. All other authors have no conflicts of interest to disclose.

SUPPLEMENTARY MATERIAL

To view supplementary material for this article, please visit https://doi.org/10.1017/S1355617720000144

REFERENCES

Alzheimer's Association. (2019). 2019 Alzheimer's disease facts and figures. *Alzheimer's Dementia*, *15*(3), 321–87. Retrieved from https://www.alz.org/facts/

Aerts, L., Heffernan, M., Kochan, N.A., Crawford, J.D., Draper, B., Trollor, J.N., Sachdev, P.S., & Brodaty, H. (2017). Effects of MCI subtype and reversion on progression to dementia in a community sample. *Neurology*, *88*(23), 2225–2232. doi: 10.1212/WNL.0000000000004015

Ahmed, S., Brennan, L., Eppig, J., Price, C.C., Lamar, M., Delano-Wood, L., Bangen, K.J., Edmonds, E.C., Clark, L., Nation, D.A., Jak, A., Au, R., Swenson, R., Bondi, M.W., & Libon, D.J. (2016). Visuoconstructional impairment in subtypes of mild cognitive impairment. *Applied Neuropsychology. Adult*, *23*(1), 43–52. doi: 10.1080/23279095.2014.1003067

Ainslie, N.K., & Murden, R.A. (1993). Effect of education on the clock drawing dementia screen in non-demented elderly persons. *Journal of the American Geriatric Society*, 1993, *41*, 249–252.

Amato, F., López, A., Peña-Méndez, E.M., Vaňhara, P., Hampl, A., & Havel, J. (2013). Artificial neural networks in medical diagnosis. *Journal of Applied Biomedicine*, *11*(2), 47–58. doi: 10.2478/v10136-012-0031-x

Autio, L., Juhola, M., & Laurikkala, J. (2007). On the neural network classification of medical data and an endeavour to balance non-uniform data sets with artificial data extension. *Computers in Biology and Medicine*, *37*(3), 388–397. doi: 10.1016/j.compbiomed.2006.05.001

Beretta, L., & Santaniello, A. (2016). Nearest neighbor imputation algorithms: a critical evaluation. *BMC Medical Informatics and Decision Making*, *16*(S3), 74. doi: 10.1186/s12911-016-0318-z

Brodaty, H., & Moore, C.M. (1997). The clock drawing test for dementia of the Alzheimer's type: a comparison of three scoring methods in a memory disorders clinic. *International Journal of Geriatric Psychiatric*, *12*, 619–627.

Brown, G. (2009). A new perspective for information theoretic feature selection. *Journal of Machine Learning Research – Proceedings Track*, *5*(1), 49–56.

Caballero, J., Ownby, R.L., Jacobs, R.J., Pandya, N., Hardigan, P.C. & Ricabal, L.C. (2018). Predicting medication adherence in older Hispanic patients with type 2 diabetes. *American Journal of Health Systems Pharmacy*, *75*(9):e194–e201. doi: 10.2146/ajhp170067.

Cahn-Weiner, D.A., Williams, K., Grace, J., Tremont, G., Westervelt, H., & Stern, R.A. (2003). Discrimination of dementia with lewy bodies from Alzheimer disease and Parkinson disease using the clock drawing test. *Cognitive and Behavioral Neurology : Official Journal of the Society for Behavioral and Cognitive Neurology*, *16*(2), 85–92.

Carew, T.G., Lamar, M., Cloud, B.S., Grossman, M., & Libon, D.J. (1997). Impairment in category fluency in ischemic vascular dementia. *Neuropsychology*, *11*(3), 400–412. doi: 10.1037/0894-4105.11.3.400

Chawla, N.V., Bowyer, K.W., Hall, L.O., & Kegelmeyer, W.P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. doi: 10.1613/jair.953

Cohen, J., Penney, D.L., Davis, R., Libon, D.J., Swenson, R.A., Ajilore, O., Kumar, A., & Lamar, M. (2014). Digital clock drawing : differentiating "Thinking" versus "Doing" in younger and older adults with depression. *Journal of the International Neuropsychological Society*, *20*, 1–9. doi: 10.1017/S1355617714000757

Cosentino, S., Jefferson, A., Chute, D.L., Kaplan, E., & Libon, D.J. (2004). Clock drawing errors in dementia: neuropsychological and neuroanatomical considerations. *Cognitive and Behavioral Neurology : Official Journal of the Society for Behavioral and Cognitive Neurology*, *17*(2), 74–84. doi: http://dx.doi.org/10.1097/01.wnn.0000119564.08162.46

Davis, R., Libon, D.J., Au, R., Pitman, D., & Penney, D.L. (2014). THink: Inferring Cognitive Status from Subtle Behaviors. In *Proceedings of the Conference on Innovative Applications of Artificial Intelligence 2014* (pp. 2898–2905). Québec City, Québec, Canada.

Davis, R., Penney, D.L., Pittman, D., Libon, D.J., Swenson, R., & Kaplan, E. (2010). The Digital Clock Drawing Test (dCDT)-I: Development of a New computerize Quantitative System. In *Annual Meeting of The International Neuropsychological Society*. Montreal, Canada.

Delis, D.C., Kramer, J., Kaplan, E., & Ober, B. (2000). *The California Berbal Learning Test-II*. New York, NY: Psychology Corporation.

Díaz-Mardomingo, M., García-Herranz, S., Rodríguez-Fernández, R., Venero, C., & Peraita, H. (2017). Problems in Classifying Mild Cognitive Impairment (MCI): One or multiple syndromes? *Brain Sciences*, *7*(12), 111. doi: 10.3390/brainsci7090111

Dykiert, D., Der, G., Starr, J.M., & Deary, I.J. (2016). Why is Mini-Mental state examination performance correlated with estimated premorbid cognitive ability? *Psychological Medicine*, *46*(12), 2647–2654. doi: 10.1017/S0033291716001045

Ehreke L., Luppa, M., Luck, T., Wiese, B., Weyerer, S., Eifflaender-Gorfer, S., Weeg, D., Olbrich, J., van den Bussche, H., Bachmann, C., Eisele, M., Maier, W., Jessen F., Wagner, M., Fuchs A., Pentzek M., Angermeyer, M.C., König H., & Riedel-Heller, S.G. (2009). Is the clock drawing test appropriate for screening for mild cognitive impairment? – Results of the German study on ageing, cognition and dementia in primary care patients (AgeCoDe). *Dementia and Geriatric Cognitive Disorders*, *28*, 365–372.

Emrani, S., Libon, D.J., Lamar, M., Price, C.C., Jefferson, A.L., Gifford, K.A., Hohman, T.J., Nation, D.A., Delano-Wood, L., Jak, A., Bangen, K.J., Bondi, M.W., Brickman, A.M., Manly, J., Swenson, R., & Au, R. (2018). Assessing working memory in Mild Cognitive Impairment with serial order recall. *Journal of Alzheimer's Disease : JAD*, *61*(3), 917–928. doi: 10.3233/JAD-170555

Emrani, S., Wasserman, V., Matusz, E., Miller, D., Lamar, M., Price, C.C., Ginsberg, T.B., Au, R., Swenson, R., & Libon, D.J. (2019). Visual versus verbal working memory in statistically-determined patients with mild cognitive impairment – On behalf of the consortium for Clinical and Epidemiological Neuropsychological Data Analysis (CENDA). *Journal of the International Neuropsychological Society, 25(10), 1001-1010.* doi: 10.1017/S1355617719000808

Eppig, J., Wambach, D., Nieves, C., Price, C.C., Lamar, M., Delano-Wood, L., Giovannetti, T., Bettcher, B.M., Penney, D.L., Swenson, R., Lippa, C., Kabasakalian, A., Bondi, M.W., & Libon, D.J. (2012). Dysexecutive functioning in mild cognitive impairment: Derailment in temporal gradients.

Journal of the International Neuropsychological Society, *18*(1), 20–28. doi: 10.1017/S1355617711001238

Feeney, J., Savva, G.M., O'Regan, C., King-Kallimanis, B., Cronin, H., & Kenny, R.A. (2016). Measurement error, reliability, and minimum detectable change in the Mini-Mental state examination, montreal cognitive assessment, and color trails test among community living middle-aged and older adults. *Journal of Alzheimer's Disease : JAD*, *53*(3), 1107–1114. doi: 10.3233/JAD-160248

Folstein, M.F., Folstein, S.E., & McHugh, P.R. (1975). "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, *12*(3), 189–198. doi: 10.1016/0022-3956(75)90026-6

Freedman, M., Leach, L., Kaplan, E., Winocur, G., Shulman, K.I., & Delis, D.C. (1994). *Clock drawing: A neuropsychological analysis*. New York, NY, US: Oxford University Press.

Hetland, A.J., Carr, D.B., Wallendorf, M.J., & Barco, P.P., (2014). Potentially driver-impairing (PDI) medication use in medically impaired adults referred for driving evaluation. *Annals of Pharmacotherapy*, *48*, 476–482.

Holsinger T., Plassman, B.L., Stechuchak, K.M., Burke, J.R., Coffman, C.J., & Williams, J.W. Jr. (2015). Stability of diagnoses of cognitive impairment, not dementia in a veterans affairs primary care population. *Journal of the American Geriatric Society*, *63*, 1105–1111.

Jak, A.J., Bondi, M.W., Delano-Wood, L., Wierenga, C., Corey-Bloom, J., Salmon, D.P., & Delis, D.C. (2009). Quantification of five neuropsychological approaches to defining mild cognitive impairment. *The American Journal of Geriatric Psychiatry : Official Journal of the American Association for Geriatric Psychiatry*, *17*(5), 368–75. doi: 10.1097/JGP.0b013e31819431d5

Kaplan, E., Goodglass, H., & Weintraub, S. (1983). *The Boston Naming Test*. Philadelphia, PA: Lea and Febiger.

Kingma, D., & Ba, J. (2014). Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

Kirby, M., Denihan, A., Bruce, I., Coakley, D., & Lawlor, B.A. (2001). The clock drawing test in primary care: sensitivity in dementia detection and specificity against normal and depressed elderly. *International Journal of Geriatric Psychiatry 16*, 935–940.

Lamar, M., Ajilore, O., Leow, A., Charlton, R., Cohen, J., GadElkarim, J., Yang, S., Zhang, A., Davis, R., Penney, D., Libon, D.J., & Kumar, A. (2016). Cognitive and connectome properties detectable through individual differences in graphomotor organization. *Neuropsychologia*, *85*, 301–309. doi: 10.1016/j.neuropsychologia.2016.03.034

Lamar, M., Price, C.C., Cynthia, D., Kaplan, E., & Libon, D.J. (2002). Capacity to maintain mental set in dementia. *Neuropsychologia*, *40*(4), 435–445. doi: 10.1016/s0028-3932(01)00125-7

Libon, D.J., Bondi, M.W., Price, C.C., Lamar, M., Eppig, J., Wambach, D.M., Nieves, C., Delano-Wood, L., Giovannetti, T., Lippa, C., Kabasakalian, A., Cosentino, S., Swenson, R., & Penney, D.L. (2011). Verbal serial list learning in mild cognitive impairment: A profile analysis of interference, forgetting, and errors. *Journal of the International Neuropsychological Society*, *17*, 905–914.

Libon, D.J., Malamut, B.L., Swenson, R., Sands, L.P., & Cloud, B.S. (1996). Further analyses of clock drawings among demented and nondemented older subjects. *Archives of Clinical Neuropsychology*, *11*(3), 193–205. doi: 10.1016/0887-6177(95)00026-7

Libon, D.J., Penney, D.L., Davis, R., Tabby, D.S., Eppig, J., Nieves, C., Bloch, A., Donohue, J.B., Brennan, L., Rife, K.L., Wicas, G., Lamar, M., Price, C.C., Au, R., Swenson, R., & Garrett, K.D. (2014). Deficits in processing speed and decision making in relapsing-remitting multiple sclerosis : the Digital Clock Drawing Test (dCDT), *Journal of Multiple Sclerosis*, *1*(2), 1–8. doi: 10.4172/jmso.1000113

McKhann, G., Drachman, D., Folstein, M., Katzman, R., Price, D., & Stadlan, E.M. (1984). Clinical diagnosis of Alzheimer's disease: Report of the NINCDS-ADRDA Work Group* under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurology*, *34*(7), 939–939. doi: 10.1212/WNL.34.7.939

Nasreddine, Z.S., Phillips, N.A., Bedirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J.L., & Chertkow, H, (2005). The Montreal Cognitive Assessment, MoCA: A brief screening tool for mild cognitive impairment. *Journal of the American Geriatric Society*, *53*, 695–699.

Pandya, S.Y., Lacritz, L.H., Weiner, M.F., Deschner, M., & Woon, F.L. (2017). Predictors of reversion from mild cognitive impairment to normal cognition. *Dementia and Geriatric Cognitive Disorders*, *43*(3–4), 204–214. doi: 10.1159/000456070

Penney, D.L., Davis, R., Libon, D.J., Lamar, M., Price, C.C., Swenson, R., Garrett, K.D., Freedland, A., Weninger, C., Scala, S., Giovannetti, T., & Kaplan, E. (2010). The Digital Clock Drawing Test (dCDT)-II: A New computerized Quantitative System. In *Annual Meeting of The International Neuropsychological Society*. Montreal, Canada.

Penney, D.L., Libon, D.J., Lamar, M., Swenson, R., Price, C.C., Weninger, C., Freedland, A., Garrett, K.D., Scala, S., & Davis, R. (2010). The Digital Clock Drawing Test (dCDT)-III: Clinician reliability for a New Quantitative System. In *Annual Meeting of The International Neuropsychological Society*. Montreal, Canada.

Price, C.C., Cunningham, H., Coronado, N., Freedland, A., Cosentino, S., Penney, D.L., Penisi, A., Bowers, D., Okun, M.S., & Libon, D.J. (2011). Clock drawing in the montreal cognitive assessment: Recommendations for dementia assessment. *Dementia and Geriatric Cognitive Disorders*, *31*(3), 179–187. doi: 10.1159/000324639

Rabin, L.A., Barr, W.B., & Burton, L.A. (2005). Assessment practices of clinical neuropsychologists in the United States and Canada : A survey of INS, NAN, and APA Division 40 members. *Archives of Clinical Neuropsychology*, *20*(October 2017), 33–65. doi: 10.1016/j.acn.2004.02.005

Reinera, K., Eichlera, T., Hertela, J., Hoffmann, W., Thyriana, J.R. (2018). The clock drawing test: A reasonable instrument to assess probable dementia in primary care? *Current Alzheimer Research*, *15*, 38–43.

Reitan, R.M., & Wolfson, D. (1985). *The Halstead-Reitan neuropsychological test battery: Theory and clinical interpretation (Vol. 4)*. Tucson, AZ: Neuropsychology Press.

Rouleau, I., Salmon, D.P., Butters, N., Kennedy, C., & McGuire, K. (1992). Quantitative and qualitative analyses of clock drawings in Alzheimer's and Huntington's disease. *Brain and Cognition*, *18*(1), 70–87.

Royall, D.R., Cordes, J.A., & Polk, M. (1998). CLOX: an executive clock drawing task. *Journal of Neurology, Neurosurgery & Psychiatry*, *64*(5), 588–594. doi: 10.1136/jnnp.64.5.588

Royall, D.R., Palmer, R.F., & Markides, K.S. (2017). Exportation and validation of latent constructs for dementia case finding in a Mexican American population-based cohort. *Journals of*

Gerontology – Series B Psychological Sciences and Social Sciences, *72*(6), 947–955. doi: 10.1093/geronb/gbw004

Schillerstrom, J.E., Sawyer Baker, P., Allman, R.M., Rungruang, B., Zamrini, E., & Royall, D.R. (2007). Clock drawing phenotypes in community-dwelling African Americans and Caucasians: Results from the University of Alabama at Birmingham study of aging. *Neuroepidemiology*, *28*(3), 175–180. doi: 10.1159/000104095

Shimada, H., Doi, T., Lee, S., & Makizako, H. (2019). Reversible predictors of reversion from mild cognitive impairment to normal cognition: A 4-year longitudinal study. *Alzheimer's Research and Therapy*, *11*(1), 1–9. doi: 10.1186/s13195-019-0480-5

Shulman, K.I. (1993). Clock-drawing and dementia in the community: A longitudinal study. *International Journal of Geriatric Psychiatry*, *8*(6), 487–496.

Souillard-Mandar, W., Davis, R., Rudin, C., Au, R., Libon, D.J., Swenson, R., Price, C.C., Lamar, M., & Penney, D.L. (2016). Learning classification models of cognitive conditions from subtle behaviors in the Digital Clock Drawing Test. *Machine Learning: Special issue for Healthcare and Medicine*, *102*, 393–441.

Spreen, O., & Strauss, E.A. (1990). *Compendium of Neuropsychological Tests*. New York: Oxford University Press.

Tuokko, H., Hadjistavropoulos, T., Miller, J.A., & Beattie, B.L. (1992). The clock test: A sensitive measure to differentiate normal elderly from those with Alzheimer disease. *Journal of the American Geriatrics Society*, *40*(6), 579–584. doi: 10.1111/j.1532-5415.1992.tb02106.x

Wasserman, V., Emrani, S., Matusz, E.F., Miller, D., Garrett, K.D., Gifford, K.A., Hohman, T.J., Jefferson, A.L., Au, R., Swenson, R., & Libon, D.J. (2019). Visual and verbal serial list learning in patients with statistically-determined mild cognitive impairment. *Innovation in Aging*, *3*(2), 1–12. doi: 10.1093/geroni/igz009

Wasserman, V., Emrani, S., Matusz, E.F., Peven, J.C., Cleary, S., Price, C.C., Ginsberg, T.B., Swenson, R., Heilman, K.M., Lamar, M., & Libon, D.J. Visuospatial Performance in Patients with Statistically-Defined Mild Cognitive Impairment (submitted for publication).

Wechsler, D. (1997). *WAIS-III: Administration and scoring manual: Wechsler adult intelligence scale*. Psychological Corporation.