Received: 26 January 2022 | Revised: 21 April 2023 | Accepted: 26 April 2023

DOI: 10.1002/dad2.12441

Alzheimer's & Dementia
Diagnosis, Assessment & Disease Monitoring

RESEARCH ARTICLE

Explainable automated evaluation of the clock drawing task for memory impairment screening

Dakota Handzlik[1] | Lauren L. Richmond[2] | Steven Skiena[1] | Melissa A. Carr[3] | Sean A. P. Clouston[4,5] | Benjamin J. Luft[3,6]

[1] Department of Computer Science, Stony Brook University, Stony Brook, New York, USA

[2] Department of Psychology, Stony Brook University, Stony Brook, New York, USA

[3] Program in Public Health, Renaissance School of Medicine, Stony Brook University, Stony Brook, New York, USA

[4] Department of Family, Population, and Preventive Medicine, Renaissance School of Medicine, Stony Brook University, Stony Brook, New York, USA

[5] World Trade Center Health and Wellness Program, Renaissance School of Medicine, Stony Brook University, Stony Brook, New York, USA

[6] Department of Medicine, Renaissance School of Medicine, Stony Brook University, Stony Brook, New York, USA

Correspondence
Sean Clouston, PhD, Program in Public Health, #3-071, Health Sciences Center, Stony Brook University, Stony Brook, NY, 11794-8338, USA.
Email: sean.clouston@stonybrookmedicine.edu

Funding information
National Institutes of Health/National Institute on Aging, Grant/Award Number: R01 AG049953

Abstract

Introduction: The clock drawing task (CDT) is frequently used to aid in detecting cognitive impairment, but current scoring techniques are time-consuming and miss relevant features, justifying the creation of an automated quantitative scoring approach.

Methods: We used computer vision methods to analyze the stored scanned images (N = 7,109), and an intelligent system was created to examine these files in a study of aging World Trade Center responders. Outcomes were CDT, Montreal Cognitive Assessment (MoCA) score, and incidence of mild cognitive impairment (MCI).

Results: The system accurately distinguished between previously scored CDTs in three CDT scoring categories: contour (accuracy = 92.2%), digits (accuracy = 89.1%), and clock hands (accuracy = 69.1%). The system reliably predicted MoCA score with CDT scores removed. Predictive analyses of the incidence of MCI at follow-up outperformed human-assigned CDT scores.

Discussion: We created an automated scoring method using scanned and stored CDTs that provided additional information that might not be considered in human scoring.

KEYWORDS

clock drawing task, Montreal Cognitive Assessment, semi-automated neurocognitive testing, World Trade Center responders

1 | INTRODUCTION

Alzheimer's disease and related dementias (ADRDs) are a major public health concern responsible for ≈1.55 million deaths and 150 million cases worldwide by 2050.[1] ADRD is characterized by the presence of cognitive impairment across multiple domains of cognition, including short-term and working memory, orientation to time and place, and/or executive functioning.[2] To facilitate diagnosis, brief cognitive assessments like the Montreal Cognitive Assessment (MoCA) have become standard clinical assessments when screening for mild cognitive impairment (MCI) and ADRD. The MoCA incorporates assessments of multiple cognitive domains and is validated for identifying prodromal ADRDs.[3] During administration of the MoCA, patients complete an analog clock drawing task (CDT) that has been used in research

RESEARCH IN CONTEXT

1. **Systematic review**: The authors reviewed the literature using standard sources as well as meeting abstracts and presentations. Although relatively few automated scoring approaches have been proposed for the clock drawing task (CDT), several recent publications present relevant approaches. We have included these citations in our submission.
2. **Interpretation**: Our study supports the use of this algorithm in settings that collect paper-and-pen CDT as part of the Montreal Cognitive Assessment (MoCA) for extracting features that may not be captured by the current MoCA-CDT scoring method. Moreover, our study provides evidence for the utility of the automated scoring algorithm described herein for understanding current cognitive status and for predicting future conversion to cognitive impairment in a large sample of World Trade Center responders in midlife.
3. **Future directions**: The application of the auto-scoring algorithm outlined here is expected to provide a deeper understanding of relevant features present in CDTs that predict current cognitive status and risk for future cognitive decline. This approach is also consistent with recent interest in characterizing neuropsychological task performance according to the process by which scores are obtained; the algorithm described here is expected to be useful in this regard.

and clinical settings for decades because of its ease of administration and sensitivity to ADRDs. Errors on the CDT can stem from a variety of underlying cognitive and motor issues including deficits in recognizing the attributes and features of a clock, visuospatial functioning, executive functioning, planning, and perseveration on clock features.[4]

Many scoring methods for the CDT have been proposed (review of proposed scoring methods in Ref 6). For MoCA-CDT scoring, a score from 0 to 3 points is assigned; prior reports of the sensitivity for detecting conversion to AD was very high (92.9%; 7).[5] However, although the CDT can be informative with trained administrators,[6] differences in training can impact scoring accuracy,[7] thereby reducing reliability[8] and face validity.[9] One study attempted to automate CDT scoring by applying machine learning (N = 1315) to raw clock images with a reported accuracy of 72.2%.[10] The approach focused on CDT screening and scoring using the six-point Shulman method, with scores ranging from perfect to severely disorganized and unrecognizable as a clock.[11] However, the clinical applications of that system were limited because the reasons behind convolutional neural network decisions were hidden, and decisions were, therefore, difficult to explain to clinical or research staff. Consequently, a digital version of the CDT (dCDT) was developed to address some of the problems identified with traditional hand-scoring approaches.[12,13] The dCDT is promising because it can distinguish healthy older adults from patients with dementia[14] and help differentiate dementia subtypes,[15] but the dCDT assessments relied on proprietary software and required that a task be completed with a digital pen that might capture drawings in real-time. The dCDT has spawned additional efforts including, for example, one study that used a proprietary CNN to determine diagnostic accuracy for MCI and ADRD (reported accuracy ranged from 83.44%–91.49%) in a study of 163 older adults.[16]

The results of previously proposed automated scoring methods for the CDT provide cause for optimism regarding the potential for automated scoring to detect subtle differences in cognitive status and to augment hand-scoring for the CDT, although wide adoption of available automated scoring methods for the CDT have stumbled at several barriers. The application of automated scoring methods to existing paper-and-pencil versions of this task might increase the utility of automated methods while reducing the costs associated with human scoring. The goal of this study was to bridge this gap by providing a method for feature extraction that offers reasonable predictive power while remaining transparent in classification decisions. This scoring method may bring practitioners' attention to new CDT features indicative of future cognitive deterioration in patients and may offer finer granularity than the binary score assigned by human annotators under the MoCA-CDT scoring method. Here, we provided a solution that may help improve screening accessibility and throughput, as well as the extraction of CDT features that might be important for understanding patients' current cognitive status and predicting future cognitive change.

2 | METHOD

2.1 | Setting

In 2014, we began to assess cognition in World Trade Center (WTC) responders by recruiting responders from a population-based monitoring program[17] and conducting an extensive functional and behavioral assessment program that included the MoCA.[18] From our previous studies we know that the prevalence of dementia and incidence of MCI has been observed at higher-than-expected rates according to the age of the cohort, with the highest rates concentrated among those who spent the longest time on site and those who developed post-traumatic stress disorder (PTSD),[19] leading many to believe that WTC responders might be at heightened risk of neurodegenerative disease.[20]

2.2 | Participants and procedure

WTC responders completed the Montreal Cognitive Assessment MoCA annually (N for the current analysis = 7109 observations among 4919 WTC responders) and overall scores and domain-specific scores were retained for domains of episodic memory, attention, abstraction, processing speed, numeracy, language, fluency, visuospatial functioning, and orientation. To facilitate long-term record storage and for quality-assurance purposes, our research protocol includes a requirement that the MoCA-CDT be scanned into a computer and stored indefinitely. Scanned images were retrieved in their native portable document format (PDF) and, for the purposes of this study, were cropped to exclude extraneous details. Scan files were labeled using a research identifier and visit date. The CDT was scored originally on a three-point scale according to the method prescribed by the MoCA scoring protocol,[3] with one point each assessed for drawing a circular clock face (contour); accurate placement, and serial order of clock numbers (digits); and hands indicating the correct time, with one hand shorter than the other and originating at the center of the clock (hands). Human-assigned scoring was done by trained staff following the MoCA's standard operating procedure, and each of the CDT scores was checked independently by a second trained research staff member to ensure consistency of scoring prior to being reported. Training for CDT scoring occurs regularly, along with regular monitoring to ensure that there is no slippage in either the delivery or scoring of the assessment. During quality control processes, CDT scores are scored independently by different trained research staff; scores that are discordant go through a third reviewer to ensure that scoring is appropriate. Records are kept of which staff members caused the error and when errors are found at a higher rate than expected at which point that staff member is retrained in both assessment and scoring. Higher scores indicate better performance for each element and for the clock drawing overall. Appendix Figure 1 has a CDT example with features highlighted and grouped according to the MoCA's predefined areas of interest (digits/contour/hands in green/red/yellow), whereas Appendix Figure 2 has an example of digits extracted from those drawings.

MCI and dementia were diagnosed algorithmically according to the diagnostic criteria recommended by the *National Institute on Aging–Alzheimer's Association*.[2,21] However, information from the clock drawing portion of the MoCA was excluded from diagnoses in this study. Diagnosis of MCI in this cohort requires a level of cognitive impairment in one or more domains that is greater than would be expected according to patient age and education level resulting from the evidence of cognitive decline at levels that do not impair activities of daily living alongside an absence of evidence of cognitive impairment sufficiently severe to significantly affect social or occupational functioning.[2] In these data, self and spousal reports of memory declines are strongly associated with psychiatric comorbidities and are thus discounted from diagnoses. Dementia requires evidence of impairment to multiple domains of cognitive functioning including in the ability to complete simple calculations, to recognize common items, and evidence of disorientation to time and place.

2.3 | Consent statement

This study was reviewed by the institutional ethics review board (#604113). Participants provided informed, written consent to participate in this study.

2.4 | Clock measures

Human-based MoCA scoring for the CDT includes gross scoring for the contour of the clock, the placement and legibility of the digits used in the clock, and the shape and direction of the hands used in the clock. As such, we group features below to similarly reflect potential sources of variability in the clock scoring and focused on including features, such as circularity of the contour, or the ratio of the length of the clock's minute and hour hands, which are commonly used to score the CDT, but also included several novel features that were considered of potential benefit.

To facilitate computerized analyses, we relied on all usable data collected over the period from January 2014 to June 2019. To be included, scans needed to be clear and needed to have been stored with appropriate meta-data so that an automated routine could reliably determine the person's research identification and the date of data collection, and so that the computer could reliably determine the location of the clock. Scans were often excluded because scans were of lower overall quality and the clock could not be seen well or was blurry, or because the scanner induced visual defects in the files. In addition, filename conventions and/or PDF-labeling standards were not always clear. In many cases, the date was not clear from either the meta-data or could not be accurately read by the computer from the PDF in a way that matched the scores. Thus, although we collected data on up to 15,298 observations, these exclusion criteria left 7109 CDTs spread at random throughout the observational period that were effectively scanned and matched to clinical data for use in the current analysis.

Automated CDT scoring focused on 19 dimensions of the Clock (Supplemental Methodological Appendix for full details). Specifically, we computed five contour measures (radius size, radius minimum/maximum ratio, circularity, center deviation, and removed points), eight measures of the digits (digit radius mean and SD, digit angle mean and SD, digit area mean and SD, as well as missing or extra digits), and finally six hand measures (intersection distance, hand angles, hands length ratio, density ratio, bound box ratio, and the number of components). These form the feature set used in the remainder of the manuscript.

2.5 | CDT auto-scoring code-sharing statement

The code that was developed for use in this project is freely accessible and can be found at https://github.com/dakota0064/WTC_Clock_Analysis.

2.6 | Validation of CDT auto-scoring

To determine the predictive power of the CDT features, as compared with the MoCA human-assigned scores, we trained a separate classifier for each of the three scoring factors. Each factor was trained by using a random forest classifier,[22] containing 100 decision trees and

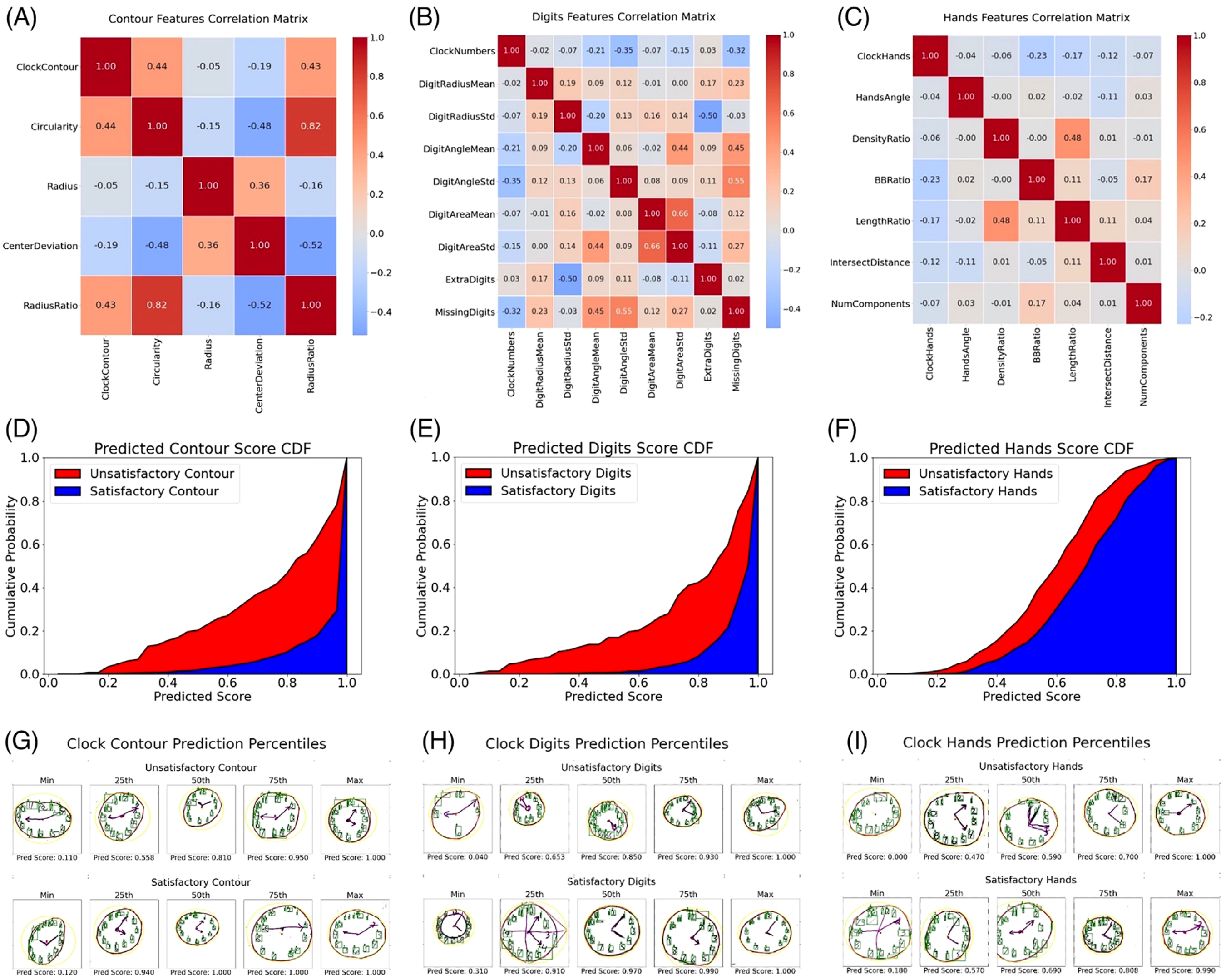

FIGURE 1 Validation data needed to examine clock scoring reliability including correlation matrices, clock feature distributions, and example clock drawings separated by human-assigned contour, digits, and hands scoring. *Note*: A–C show correlation matrices comparing contour features and human-assigned scoring for (A) contours, (B) digits, and (C) hand scores; darker red identifies stronger associations that are more strongly positive, while deeper blue shows inverse associations. D–F show the cumulative distribution function for predicted (D) contour, (E) digits, and (F) hand scores; the two curves correspond to cohorts defined by the ground truth score assigned to the clock contour by human annotators (unsatisfactory = 0, and satisfactory = 1) and the area between curves shows reasonable separation between classes. G–I show annotated clock drawings representing predicted score percentiles in the lowest quartile, median, and top quartile for each (G) contour, (H) digits, and (I) hand human-assigned scores; green labels the digits identified by the computer, purple identifies the hands, and yellow shows the circle that best fits the centroid.

no pre-set stopping depth. The models were trained on 5400 images and evaluated on a holdout set of 1422 images. Each model provided an overall forest score matched by feature type with the human-assigned MoCA-CDT scoring. For this purpose, we reported overall accuracy as well as the cumulative distribution function (CDF) showing the score distribution matching the forest scores to the MoCA-CDT scores.

We validated results through analyses showing the matches between our feature description metrics and overall human-assigned MoCA-CDT scores (range: 0–3). To elucidate the sources of the combined contributing features of each subdomain to the clock contours, we used logistic regression to examine multivariable predictive associations between the computer-assigned MoCA-CDT scores with human-assigned CDT scores. To understand how the extracted features related to performance in a variety of cognitive domains and with cognitive impairment, we used age-adjusted Spearman's rank correlation coefficients to examine associations between extracted features and neurocognitive test scores. In addition, we report demographics, including age in years and gender. Finally, for sensitivity analyses we retrieved both the clinical diagnosis and current symptomatology for PTSD and major depression.

We finished by examining the CDT's ability to predict incident MCI on follow-up testing occasions, occurring approximately annually (e.g., at least 366 days after baseline assessment). We assessed the incidence of MCI at follow-up among individuals who were cognitively

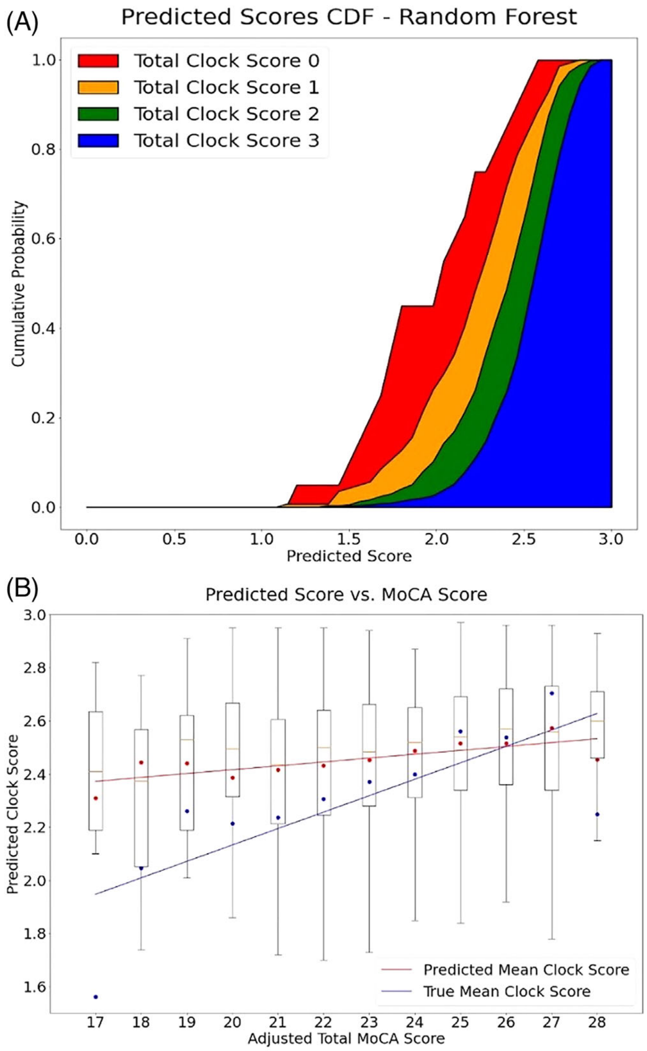

FIGURE 2 Validation data needed to examine overall clock drawing task. *Note*: A shows the cumulative distribution function plot for overall clock drawing task scores. Ground truth scores defined by the total score received on the clock drawing task as determined by human annotators: total score of 0 ($n = 20$), total score of 1 ($n = 141$), total score of 2 ($n = 532$), and total score of 3 ($n = 729$). The curves demonstrate the expected separation, although all lower scores show a significant skew upward, probably because of the general data imbalance. B shows the box and whisker plot of the summed predicted clock scores compared with the adjusted total Montreal Cognitive Assessment (MoCA) score. Adjusted total MoCA score is provided by subtracting the human-assigned clock score from the original total score. Regression lines of the mean predicted score (red) and the mean ground truth scores (blue) for each MoCA score group are shown superimposed over the plot.

unimpaired at baseline by using Cox proportional hazard modeling to estimate the years since follow-up until the incidence of MCI. The Efron method was used to handle ties; participants were censored at the time of the final follow-up examination or the date when data were retrieved. Models included both the human-assigned and feature-derived autoscoring for each component (hands, contour, and digits), and were adjusted for demographics at baseline.

TABLE 1 Characteristics describing the World Trade Center responder sample.

Baseline characteristics ($N = 4919$)	Mean	SD
Age, years	54.02	8.35
Global cognition	25.78	2.58
Cognitive performance		
Episodic memory	8.56	1.42
Processing speed	0.12	0.05
Numeracy	4.28	1.10
Language	1.60	0.61
Fluency	2.91	0.30
Orientation	5.96	0.20
Visuospatial ability	1.27	0.64
Attention	2.84	0.40
	%	
Female, %	10.23	
Educational attainment		
High school or less	26.12	
Some college	44.88	
College degree	29.00	
Race/Ethnicity		
White	82.80	
Black	5.78	
Other	4.11	
Hispanic	7.31	
Status at assessment		
Mild cognitive impairment	16.19	
Dementia	1.72	
Incidence rate at follow-up ($N = 2190$)	IR	95% CI
Mild cognitive impairment	11.94	11.38–12.53
Dementia	1.64	1.46–1.85

Abbreviations: IR, incidence rate; 95% CI, 95% confidence interval.

3 | RESULTS

The average participant in this study was in their mid to late fifties (Table 1) and the majority were men. Incidence of MCI and dementia was relatively high in this population at follow-up. Nearly all (99%) were employed full time as responders during the events at the WTC on 9/11/2001, so this variable has not been reported. The average period between observations was 1.51 (SD = 1.36) years.

3.1 | Development and initial validation of auto-scoring methodology

The contour scores determined by human annotators showed a significant skew. Within our test set, 8.9%, 11.4%, and 36.4% of the scores

TABLE 2 Breakdown of the test set into various cohorts determined by human-assigned scoring on each of the categories.

Cohort	Size	Mean Montreal Cognitive Assessment score	Mean predicted category score	Pearson	*p*
Contour Score = 0	148	24.135	0.734	0.022	0.787
Contour Score = 1	1274	25.779	0.940	0.051	0.069
Digits Score = 0	154	23.208	0.755	0.043	0.596
Digit Score = 1	1268	25.900	0.930	0.130	<**0.001**
Hands Score = 0	572	24.318	0.583	0.088	**0.035**
Hands Score = 1	850	26.477	0.679	0.037	0.277
Total Clock Score = 0	20	21.050	2.010	−0.131	0.582
Total Clock Score = 1	141	23.241	2.242	0.048	0.571
Total Clock Score = 2	532	24.885	2.412	0.111	**0.010**
Total Clock Score = 3	729	26.719	2.569	0.055	0.141

were 0 in the contour, digit, and hand scores, respectively. Figure 1A–C shows intercorrelations between various auto-scoring features and human-assigned scoring. These revealed, for example, that circularity and the radius-ratio were moderately associated with human-assigned contour scores. A classifier using the features Circularity and Radius-Ratio achieved 92.2% accuracy on the test contour set. Figure 1D–F shows CDF curves for both human-assigned scoring groups by score type and revealed that most clocks with predicted scores $\leq$0.9 had unsatisfactory contour ratings. Figure 1G–I shows representative clock examples for each human-assigned label. In general, CDTs rated as satisfactory by human hands had evidence of human error, whereas variation in the 75th percentile group revealed that humans missed variability captured by automated scores.

Scoring suggested that the distributions of aggregate predicted scores for each of the four possible total scores on the MoCA-CDT were similar (Figure 2A). The box-and-whisker plot (of predicted scores vs the total MoCA scores (Figure 2B) suggested that CDT scores could not completely replicate overall MoCA score without information from other cognitive domains.

Table 2 shows a breakdown of the mean MoCA scores and predicted overall CDT scores by feature score. Significant associations were observed between predicted overall CDT and mean MoCA scores for drawings receiving scores indicating digits = 1, hand = 0, or clock = 2. These values indicate that our system achieved finer-grained predictive resolution when CDT performance was used as a proxy for overall MoCA performance.

To determine which features were driving annotator scores, logistic regression was used to predict human-assigned binary scoring in each category (Table 3). Both the circularity and radius ratio significantly predicted the clock contour scores, whereas the radius and center deviation did not. For clock digits, features associated with the digit radius and digit angle, as well as missing digits, significantly predicted the overall clock digit scores. All other clock digit features were non-significantly predictive of overall clock digit scores. Finally, clock hand scores were associated with all relevant features except for number of components.

Next, we examined how extracted features related to performance across cognitive domains (Table 4) and found that across all categories poorer clock scores were associated with older age, slower processing speed, and higher intra-individual variability. Poorer contour scores were associated with poorer attention scores, whereas poorer digit and hand scores were associated with poorer visual memory. Poorer hand and digit scores were associated with higher atrophy risk.

To assess the clinical utility of extracted scores we examined the power to predict MCI at follow-up (Table 5). In these results, we observed that computer-defined features predicted the incidence of MCI in a jointly estimated model and after adjusting for demographics. The human-assigned clock hands score was also statistically significant in both models (Table 4). Together, these results indicated the clinical relevance of the feature extraction and auto-scoring method applied herein, as it relates to both current cognitive status and prediction of future conversion to MCI at follow-up.

3.2 | Sensitivity analyses

Chronic psychiatric disorders are common in traumatized groups, so we examined the sensitivity of associations between the human- and computer-rated clock scores in responders with chronic PTSD or those with chronic depression, as compared to those without any signs of either, to examine the sensitivity of scoring to psychiatric disorders. Overall, results suggested that the correlations were similar across populations with and without psychiatric disorders. For example, rho in patients with diagnoses of chronic PTSD (PTSD checklist score < 30, n = 162; digits = 0.81; hands = 0.92; and contour = 0.83) was slightly lower but not substantively different from rho in patients who never had a diagnosis and did not report symptoms of PTSD (PTSD checklist score < 30, n = 4,179; digits = 0.91; hands = 0.92; and contour = 0.89).

TABLE 3 Multivariable associations derived from logistic regression examining the associations between qualitative scores assigned by research staff as compared to both measured clock features and overall scores derived using random forest modeling.

	Clock contour			Clock digits			Clock hands		
Feature	**B**	**SE**	***p***	**B**	**SE**	***p***	**B**	**SE**	***p***
Circularity	17.671	5.141	**0.001**						
Radius ratio	17.217	1.396	<**0.001**						
Radius	0.000	0.002	0.868						
Center deviation	0.638	12.264	0.958						
Digit radius, mean				2.700	1.240	0.029			
Digit radius, SD				-13.346	2.865	<**0.001**			
Digit angle, mean				-2.408	0.975	**0.014**			
Digit angle, SD				-20.533	1.997	<**0.001**			
Digit area, mean				-1.720	1.104	0.119			
Digit area, SD				-0.935	0.592	0.114			
Extra digits				0.968	0.685	0.158			
Missing digits				-3.921	0.470	<0.001			
Angle of hands							-0.767	0.142	<**0.001**
Density ratio							-0.428	0.198	**0.030**
BB ratio							-3.345	0.220	<**0.001**
Length ratio							-1.856	0.203	<**0.001**
Distance from intersection							-2.413	0.294	<**0.001**
Number of components							-0.427	0.347	0.219
Feature score									
Forest contour	11.156	0.360	<**0.001**	-0.809	0.514	0.116	-0.227	0.303	0.454
Forest digits	0.132	0.308	0.667	0.699	0.320	**0.029**	11.477	0.266	<**0.001**
Forest hands	0.540	0.424	0.203	13.22	0.517	<**0.001**	1.7496	0.283	<**0.001**

Note: Coefficients (B) and standard error (SE) estimated using logistic regression. Results reported in bold typeface when they were statistically significant after adjusting for the false discovery rate.

TABLE 4 Age-adjusted Spearman's correlation coefficients (rho) showing level of association between clock features measured using the human-assigned scores and random-forest–derived automated feature scores and age as well as cognitive functioning overall and separated into nine functional subdomains.

	Clock contour	Clock digits	Clock hands	Forest contour	Forest digits	Forest hands
	Rho	**Rho**	**Rho**	**Rho**	**Rho**	**Rho**
Age, years	0.039	0.026	0.030	0.037	0.021	0.027
Montreal Cognitive Assessment without clock scoring	0.062	0.149	0.182	0.078	0.154	0.168
Abstraction	-0.010	0.031	0.045	-0.012	0.036	0.039
Processing speed	0.046	0.074	0.076	0.041	0.072	0.076
Numeracy	0.031	0.082	0.108	0.035	0.089	0.094
Language	0.014	0.030	0.021	0.024	0.029	0.018
Fluency	0.018	0.024	0.055	0.018	0.023	0.054
Orientation	0.029	0.071	0.032	0.032	0.055	0.031
Visuospatial function	0.081	0.115	0.170	0.084	0.125	0.159
Attention	0.024	0.067	0.064	0.029	0.060	0.055
Episodic memory	0.044	0.079	0.113	0.053	0.092	0.105

Note: All scores have been adjusted so that higher scores indicate poorer performance across all measures. Intensity of the red shading indicates the significance level of each association, with cells reporting no significant associations in white, significant but overall weaker associations in pink and lighter red colors, and stronger and statistically significant associations in dark red colors.

TABLE 5 Comparison of association between computer-assigned feature composite scores and human-assigned scoring methods with contemporary presence and later incidence of mild cognitive impairment.

	Prevalent mild cognitive impairment contemporaneous with scoring			Incident mild cognitive impairment after scoring		
Feature	aOR	95% CI	*p*	aHR	95% CI	*p*
Human-scored clock hands	**0.62**	**0.48–0.80**	<0.001	**0.38**	**0.28–0.51**	<0.001
Forest clock hands	0.70	0.47–1.04	0.075	**0.54**	**0.33–0.88**	**0.013**
Human-scored contour	1.07	0.70-1.64	0.754	1.09	0.73–1.64	0.673
Forest contour	**0.45**	**0.25–0.82**	**0.009**	**0.48**	**0.26–0.88**	**0.017**
Human-scored digits	0.74	0.48–1.13	0.164	0.84	0.56–1.25	0.395
Forest digits	**0.33**	**0.17–0.61**	<0.001	**0.26**	**0.14–0.47**	<0.001

Note: aOR, multivariable-adjusted odds ratio; aHR, multivariable-adjusted hazards ratio; 95% CI, 95% confidence interval; P, nominal *p*-value derived from a Cox proportional hazards regression model. All models shown additionally adjust for age in years, sex/gender, and race/ethnicity. Results reported in bold typeface when they were statistically significant after adjusting for the false discovery rate.

4 | DISCUSSION

The CDT is a clinical tool that by itself can be used to determine the presence and severity of cognitive impairment but is also included as a sub-task of the MoCA. The goal of the present work was to determine whether an automated feature extractor could reliably replicate or improve upon the human-assigned MoCA-CDT scoring. Results suggested that the computer-derived feature scores supported human-assigned scores and MoCA-CDT scores, but CDT scores derived from the machine-learning algorithm were not able to reliably replicate the full MoCA score. Results suggest that the auto-scoring system achieved finer-grained predictive resolution when CDT performance was used as a proxy for overall MoCA performance. The study also suggested that several features identified by the machine-learning algorithm provided information that was ignored by human raters. These improvements appeared to emerge from the confluence of smaller differences, which although evident to the computer were overlooked by raters. Results support the use of computer vision and machine-learning methods to replace human CDT scoring.

Turning to validation of our machine-learning algorithm results overall speak to the clinical relevance of this scoring method as related to both current cognitive status and the prediction of future conversion to MCI at follow-up. Since 2014, the MoCA-CDT has been collected as part of a study of responders who worked at the WTC sites after the tragic events on 9/11/2001, and we have included the scoring provided by human raters in epidemiological analyses focused on understanding brain health.[23] Results suggest that the auto-scoring method could be useful for distinguishing patients who are experiencing clinically relevant cognitive changes indicative of higher risk for MCI compared to those who might remain cognitively normal. Specifically, auto-scoring predicted future conversion to MCI for participants who were deemed to be cognitively normal on initial CDT assessments, whereas human scoring predicted only the incidence of MCI when examining clock hands. Together, findings support the use of the auto-scoring method for detecting subtle distinctions between participants who are cognitively normal at baseline and develop MCI versus those who remain cognitively normal. The ability to distinguish two different subgroups is important, since hand-scoring methods for the CDT have low sensitivity for distinguishing MCI and mild AD.[24]

One advantage of the current approach over past auto-scoring approaches is that the features extracted by the current machine-learning algorithm can be described in terms that are easily understandable to human scorers (e.g., “missing digits” and “extra digits” under the digit feature). This aspect might enable a more complete description of how participants received a specific score; for example, one participant might receive a poor digit score because of missing digits, whereas another participant might receive the same score by adding extra digits. There has been renewed interest in characterizing neuropsychological test performance using process-based approaches, particularly as mediated through the technology.[25] In contrast to previously proposed “black box” approaches to auto-scoring for the CDT,[10] CDT features extracted through the current approach may improve the ability to explain results.

Future studies may adopt the auto-scoring method described herein to validate this approach in a greater variety of psychiatric and neurological populations, and to conduct finer-grained analyses of the extracted features' relationship with cognitive status. For example, in studies of action execution, action omission has been associated with amnestic MCI, whereas commission errors (including repeated actions) have been associated with dysexecutive MCI.[26] Similar patterns emerge for missing digits (omissions) and extra digits (commissions) in patients with varying degrees and types of cognitive impairment than were tested in the current study.

5 | LIMITATIONS

Despite being the largest study to date to examine digitized images of hand-drawn versions of the CDT completed during in-person administration of the MoCA, this study had several limitations. First, individuals were assessed at midlife when neurodegenerative diseases are relatively uncommon. This might have improved the training of the method to recognize variation seen in normal functioning and improved sensitivity for detecting early forms of cognitive decline but may have also limited our ability to determine features that might be specific to more severe disease or to different disease subtypes. Moreover, this study focused on a cohort of WTC responders mostly comprising men who are thought to be at heightened risk of brain aging due to their exposures on site during the response efforts. Consequently, this work may benefit from replication efforts in other populations.

The current approach tested model performance of a computer-generated CDT scoring metric against human CDT scoring provided by trained raters as the "ground truth." Given the issues with human-assigned scoring, human scoring measures do not appear to be optimal for evaluating model performance. Future efforts may consider raters' number of years of experience in scoring the MoCA-CDT and/or average scores awarded by specific raters to account for inter-rater differences to attempt to control for idiosyncrasies that might arise from measures relying on hand-scoring by different raters.

Although prior studies of CDT have been conducted in samples with mixed MCI/dementia, the WTC sample is a population cohort and is composed primarily of individuals with normal cognitive functioning. Therefore, when compared to clinical cohorts, the WTC responder sample reported herein might be expected to exhibit better performance on the MoCA and particularly on the CDT, given the relative preservation of cognitive function in this sample. The potential for the WTC cohort to score better on the CDT than clinical studies have two practical consequences for the current investigation. First, scores clustered at the higher end of the range, rather than more evenly distributed across the range, may limit the extent to which the current automated scoring methodology could be applied to clock drawings produced by more impaired populations. Concurrently, development of the automated scoring method in a relatively less-impaired sample compared to those used in prior studies may be useful for identifying CDT features that can reliably differentiate clinically relevant MCI from normal cognitive aging. Distinguishing normal cognitive aging from MCI can be difficult to do in clinical settings but is quite important from the standpoint of the potential for early intervention and behavioral management of symptomology that could be especially helpful for individuals with MCI.[27]

5.1 | Clinical implications

This large-scale study demonstrated that an automated scoring for the CDT could be used with standard administration techniques in a way that improved test sensitivity. CDT auto-scoring methods provide an objective score for the CDT, while reducing the burden of human scoring the CDT. The algorithm allows for a characterization of CDT performance, which may be able to differentiate normal cognitive functioning more sensitively from MCI in mid to late life. In addition, it generates features such as clock size and pen pressure that may be sensitive to mobility or behavioral differences. Finally, our auto-scoring method opens the door to the use of open-source electronic tools for collecting CDTs in clinical settings and could allow for the use of phone apps or other electronic means of producing CDTs to facilitate patient monitoring. Together, these results support the clinical utility of CDT automated scoring methods when characterizing MCI and prodromal ADRD.

ACKNOWLEDGMENTS

The authors would like to acknowledge funding to examine aging in World Trade Center responders provided by the National Institutes of Health/National Institute on Aging (NIH/NIA R01 AG049953).

CONFLICT OF INTEREST STATEMENT

No conflicts of interest to disclose.

CONSENT STATEMENT

All human subjects provided written consent.

ORCID

Lauren L. Richmond https://orcid.org/0000-0001-6793-5773
Sean A. P. Clouston https://orcid.org/0000-0002-6124-0329

REFERENCES

1. Nichols E, Vos T. The estimation of the global prevalence of dementia from 1990-2019 and forecasted prevalence through 2050: An analysis for the Global Burden of Disease (GBD) study 2019. *Alzheimer Dement*. 2021;17:e051496.
2. McKhann GM, Knopman DS, Chertkow H, et al. The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer Dement*. 2011;7:263-269.
3. Nasreddine ZS, Phillips NA, Bedirian V, et al. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. *J Am Geriatr Soc*. 2005;53:695-699.
4. Eknoyan D, Hurley RA, Taber KH. The clock drawing task: common errors and functional neuroanatomy. *J Neuropsychiatry Clin Neurosci*. 2012;24:260-265.
5. Wang P, Shi L, Zhao Q, Hong Z, Guo Q. Longitudinal changes in clock drawing test (CDT) performance before and after cognitive decline. *PLoS One*. 2014;9:e97873.
6. Price CC, Cunningham H, Coronado N, et al. Clock drawing in the Montreal Cognitive Assessment: recommendations for dementia assessment. *Dement Geriatr Cogn Disord*. 2011;31:179-187.
7. Frei BW, Woodward KT, Zhang MY, et al. Considerations for clock drawing scoring systems in perioperative anesthesia settings. *Anesth Analg*. 2019;128:e61.
8. Feeney J, Savva GM, O'Regan C, King-Kallimanis B, Cronin H, Kenny RA. 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. *J Alzheimer Dis*. 2016;53:1107-1114.
9. Seigerschmidt E, Mösch E, Siemen M, Förstl H, Bickel H. The clock drawing test and questionable dementia: reliability and validity. *Int J Geriatr Psychiatry*. 2002;17:1048-1054.
10. Chen S, Stromer D, Alabdalrahim HA, Schwab S, Weih M, Maier A. Automatic dementia screening and scoring by applying deep learning on clock-drawing tests. *Sci Rep*. 2020;10:1-11.
11. Shulman KI, Shedletsky R, Silver IL. The challenge of time: clock-drawing and cognitive function in the elderly. *Int J Geriatr Psychiatry*. 1986;1:135-140.

12. Davis R, Penney D, Pittman D, Libon D, Swenson R, Kaplan E. *The Digital Clock Drawing Test (dCDT) I: Development of a New Computerized Quantitative System*. The International Neuropsychological Society; 2011.
13. Penney D, Davis R, Libon D, et al. *The digital clock drawing test (DCDT)-II: A New Computerized Quantitative System*. Annual Meeting of The International Neuropsychological Society; 2010.
14. Davis R, Libon D, Au R, Pitman D, Penney D. THink: inferring cognitive status from subtle behaviors. *AI Magazine*. 2015;36:49-60.
15. Souillard-Mandar W, Davis R, Rudin C, et al. Learning classification models of cognitive conditions from subtle behaviors in the digital clock drawing test. *Machine Learning*. 2016;102:393-441.
16. Binaco R, Calzaretto N, Epifano J, et al. Machine learning analysis of digital clock drawing test performance for differential classification of mild cognitive impairment subtypes versus Alzheimer's disease. *J Int Neuropsychol Soc*. 2020;26:690-700.
17. Dasaro CR, Holden WL, Berman KD, et al. Cohort profile: world trade center health program general responder cohort. *Int J Epidemiol*. 2017;46:e9.
18. Clouston SA, Diminich ED, Kotov R, et al. Incidence of mild cognitive impairment in World Trade Center responders: long-term consequences of re-experiencing the events on 9/11/2001. *Alzheimer Dement*. 2019;11:628-636.
19. Clouston SAP, Hall CB, Kritikos M, et al. Cognitive impairment and World Trade Centre-related exposures. *Nat Rev Neurol*. 2022;18:103-116.
20. Daniels RD, Clouston SAP, Hall CB, et al. A workshop on cognitive aging and impairment in the 9/11-Exposed Population. *Int J Environ Res Public Health*. 2021;18:681.
21. Albert MS, DeKosky ST, Dickson D, et al. The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer Dement*. 2011;7:270-279.
22. Pal M. Random forest classifier for remote sensing classification. *Int J Remote Sens*. 2005;26:217-222.
23. Mukherjee S, Clouston S, Kotov R, Bromet E, Luft B. Handgrip Strength of World Trade Center (WTC) responders: the Role of Re-Experiencing Posttraumatic Stress Disorder (PTSD) symptoms. *Int J Environ Res Public Health*. 2019;16:1128.
24. Powlishta K, Von Dras D, Stanford A, et al. The clock drawing test is a poor screen for very mild dementia. *Neurology*. 2002;59:898-903.
25. Diaz-Orueta U, Blanco-Campal A, Lamar M, Libon DJ, Burke T. Marrying past and present neuropsychology: is the future of the process-based approach technology-based? *Front Psychol*. 2020;11:361.
26. Giovannetti T, Bettcher BM, Brennan L, Libron DJ, Kessler RK, Duey K. Coffee with jelly or unbuttered toast: commissions and omissions are dissociable aspects of everyday action impairment in Alzheimer's disease. *Neuropsychology*. 2008;22:235.
27. Jekel K, Damian M, Wattmo C, et al. Mild cognitive impairment and deficits in instrumental activities of daily living: a systematic review. *Alzheimer Res Therapy*. 2015;7:1-20.

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: Handzlik D, Richmond LL, Skiena S, Carr MA, Clouston SAP, Luft BJ. Explainable automated evaluation of the clock drawing task for memory impairment screening. *Alzheimer's Dement*. 2023;15:e12441. https://doi.org/10.1002/dad2.12441