THink: Inferring Cognitive Status from Subtle Behaviors

Randall Davis, David J. Libon, Rhoda Au,
David Pitman, Dana L. Penney

■ *The digital clock drawing test is a fielded application that provides a major advance over existing neuropsychological testing technology. It captures and analyzes high-precision information about both outcome and process, opening up the possibility of detecting subtle cognitive impairment even when test results appear superficially normal. We describe the design and development of the test, document the role of AI in its capabilities, and report on its use over the past seven years. We outline its potential implications for earlier detection and treatment of neurological disorders. We set the work in the larger context of the THink project, which is exploring multiple approaches to determining cognitive status through the detection and analysis of subtle behaviors.*

We are generally accustomed to assessment of our physical health (for example, exams that include measuring blood pressure, cholesterol, and others), but perhaps less familiar with the concept of assessing cognitive health, that is, determining the condition of the wide variety of our cognitive capabilities. *Neurocognitive testing* is the overall term for the efforts to assess the performance of our mental capabilities, including for example, memory, attention, problem solving, language and verbal fluency, cognitive processing speed, and others. Cognitive assessment is important in a variety of medical circumstances, including head injury (for example, concussions occurring during sports or on the battlefield), stroke, the onset of declining cognitive capacity resulting from dementia (for example, Alzheimer's), and others. Routine cognitive assessment has recently been added to Medicare's annual wellness visits.

We describe a new means of doing neurocognitive testing, enabled through the use of an off-the-shelf digitizing ballpoint pen from Anoto, Inc., combined with novel software we have created. The new approach provides a significant increase in precision, improves efficiency, sharply reduces test analysis time, and permits administration and analysis of the test by medical office staff (rather than requiring time from clinicians). In addition, where previous approaches to test analysis involve criteria phrased in qualitative terms, leaving room for differing interpretations, our analysis routines are embodied in code, reducing the chance for subjective judgments and measurement errors.

The digital pen provides data two orders of magnitude more precise than pragmatically available previously, making it possible for our software to detect and measure new phenomena. Because the data provide timing information, our test measures elements of cognitive processing, allowing us for example to calibrate the amount of effort subjects are expending, independent of whether their final results appear normal. As we discuss below, this has interesting implications for detecting and treating impairment before it manifests clinically.

The Clock Drawing Test

For more than 50 years clinicians have been giving the clock drawing test, a deceptively simple yet widely accepted cognitive screening test able to detect altered cognition in a wide range of neurological disorders, including dementias (for example, Alzheimer's), stroke, Parkinson's, and others (Freedman et al. 1994, Grande et al. 2013). The test instructs the subject to draw on a blank page a clock showing 10 minutes after 11, then asks the subject to copy a predrawn clock showing that time. The two parts of the test are purposely designed to test differing aspects of cognition: the first challenges things like language and memory, while the second tests aspects of spatial planning and executive function (the ability to plan and organize).

As widely accepted as the test is, there are drawbacks, including variability in scoring and analysis, and reliance on either a clinician's subjective judgment of broad qualitative properties (Nair et al. 2010) or the use of a labor-intensive evaluation system. One scoring technique, for instance, calls for appraising the drawing by eye, giving it a 0–3 score, based on measures like whether the clock circle has "only minor distortion," whether the hour hand is "clearly shorter" than the minute hand, and others, without ever defining these criteria clearly (Nasreddine et al. 2005). More complex scoring systems (for example, Nyborn et al. [2013]) provide more information, but may require significant manual labor (for example, use of rulers and protractors), and are as a result far too labor intensive for routine use.

The test is used across a very wide range of ages — from the 20s to well into the 90s — and cognitive status, from healthy to severely impaired cognitively (for example, Alzheimer's) or physically (for example, tremor, Parkinson's). Clocks produced may appear normal (figure 1a) or be quite blatantly impaired, with elements that are distorted (figure 1b), misplaced (figure 1c), repeated, missing entirely, or incomprehensible because they are overwritten (for example, figure 1d). As we explore below, clocks drawn by clinically healthy people are not always free of error, while static clock drawings that look normal on paper may have evidence of impairment evident in the drawing process.

Our System

Since 2006 we have been administering the test using a digitizing ballpoint pen from Anoto, Inc.[1] The pen functions in the subject's hand as an ordinary ballpoint, but simultaneously measures its position on the page every 12 milliseconds with an accuracy of ±0.002 inches. We refer to the combination of the pen data and our software as the digital clock drawing test (dCDT); it is one of several innovative tests being explored by the THink project.

Our software is device independent in the sense that it deals with time-stamped data and is agnostic about the device. We use the digitizing ballpoint because a fundamental premise of the clock drawing test is that it captures the subject's normal, spontaneous behavior. Our experience is that subjects accept the digitizing pen as simply a (slightly fatter) ballpoint, unlike tablet-based tests, about which subjects sometimes express concern. Use of a tablet and stylus may also distort results by its different ergonomics and its novelty, particularly for older subjects or individuals in developing countries. While not inexpensive, the digitizing ballpoint is still more economical, smaller, and more portable than current handheld devices, and is easily shared by staff, facilitating use in remote and low-income populations.

The dCDT software we developed is designed to be useful both for the practicing clinician and as a research tool. The program provides both traditional outcome measures (for example, are all the numbers present, and in roughly the right places) and, as we discuss below, detects subtle behaviors that reveal novel cognitive processes underlying the performance.

Figure 2 shows the system's interface. Basic information about the subject is entered in the left panel (anonymized here); folders on the right show how individual pen strokes have been classified (for example, as a specific digit, hand, and others.). Data from the pen arrives as a set of strokes, composed in turn of time-stamped coordinates; the center panel can show that data from either or both of the drawings, and is zoomable to permit extremely detailed visual examination of the data if needed.

In response to pressing one of the classify buttons, the system attempts to classify each pen stroke in a drawing. Colored overlays are added to the drawing (figure 3, a close-up view) to make clear the resulting classifications: tan bounding boxes are drawn around each digit, an orange line shows an ellipse fit to the clock circle stroke(s), green and purple highlights mark the hour and minute hands.

The classification of the clock in figure 3 is almost entirely correct; the lone error is the top of the 5, drawn sufficiently far from the base that the system

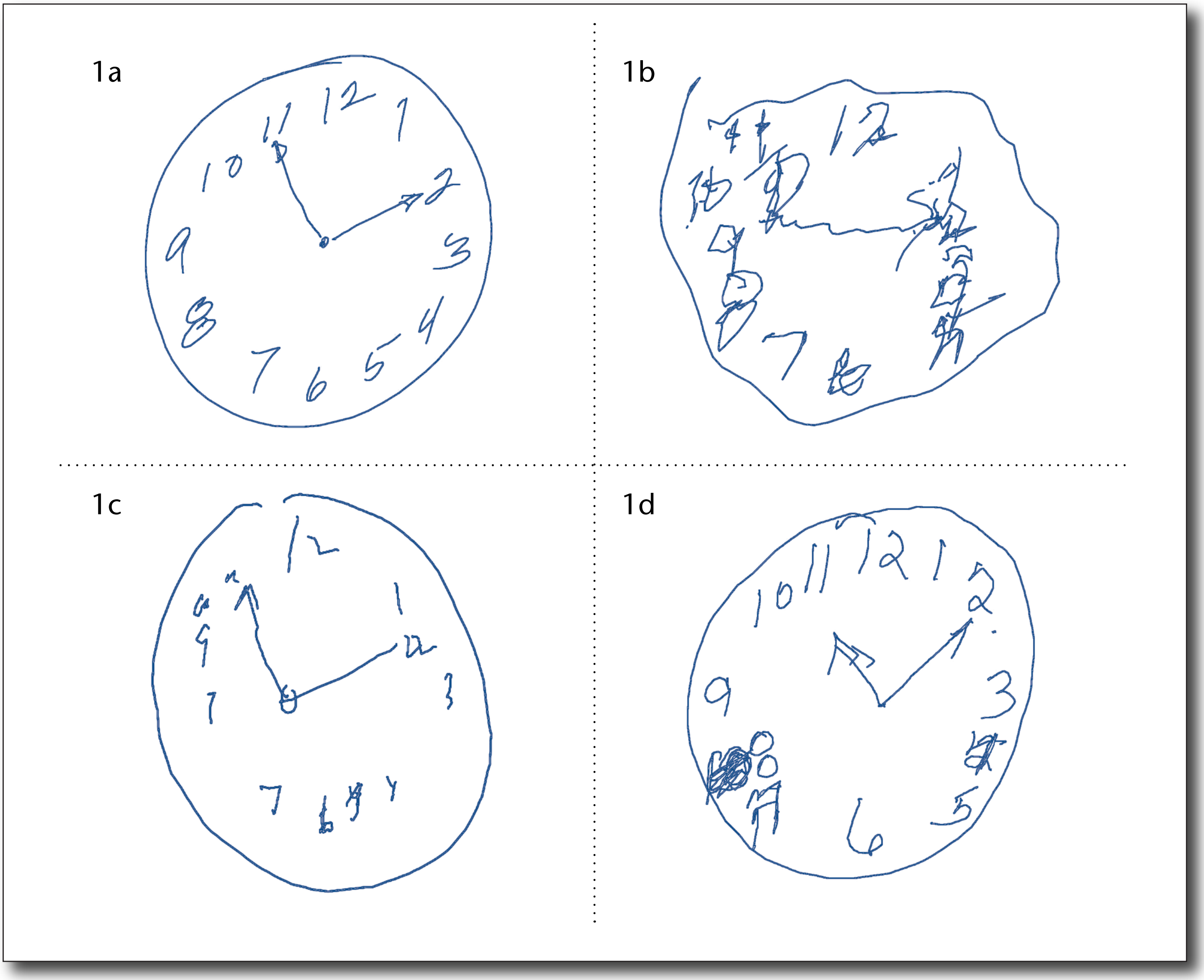
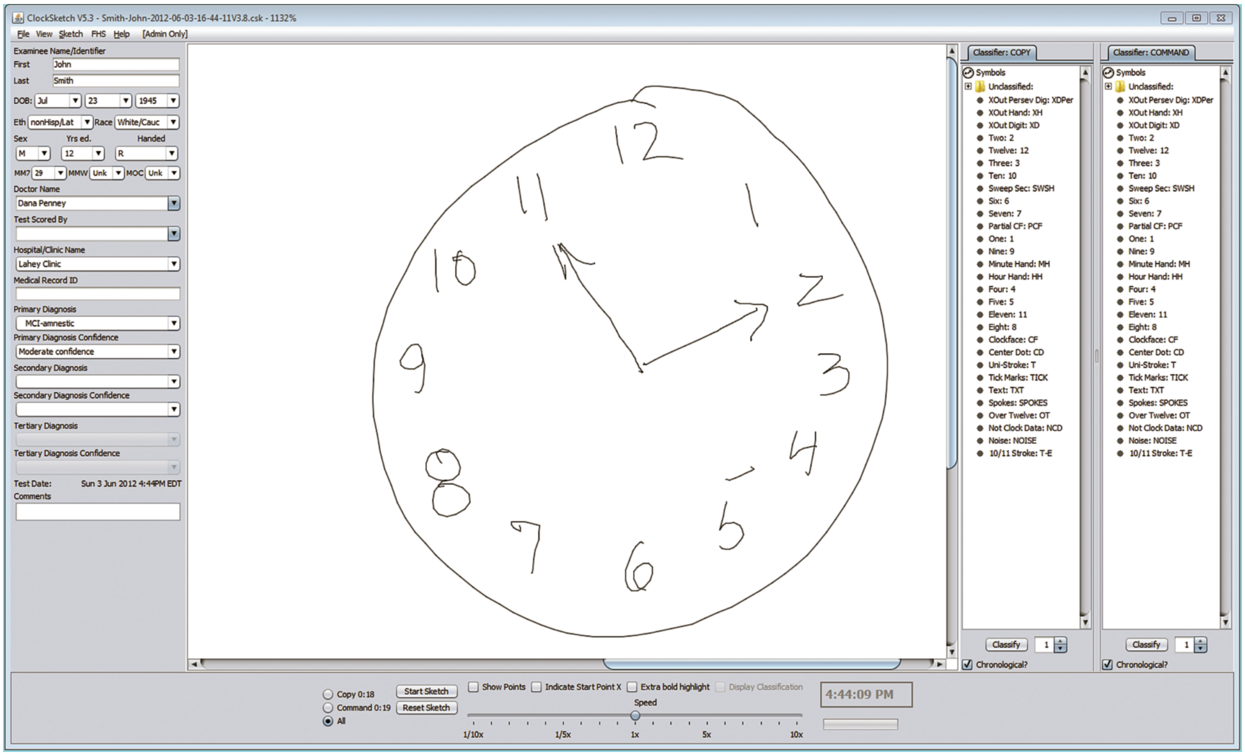


Figure. 1. Example Clocks.

Normal appearing (1a), clearly impaired (1b, 1c), overwritten (1d).

missed it. The system has put the stroke into a folder labeled simply “Line.” The error is easily corrected by dragging and dropping the errant stroke into the “Five” folder; the display immediately updates to show the revised classification. Given the system’s initial attempt at classification, clocks from healthy or only mildly impaired subjects can often be correctly classified in 1–2 minutes; unraveling the complexities in more challenging clocks can take additional time, with most completed within 5 minutes. The drag and drop character of the interface makes classifying strokes a task accessible to medical office staff, freeing up clinician time. The speedy updating of the interface in response to moving strokes provides a gamelike feeling to the scoring that makes it a reasonably pleasant task.

Because the data is time stamped, we capture both the end result (the drawing) and the behavior that produced it: every pause, hesitation, and time spent simply holding the pen and (presumably) thinking, are all recorded with 12 millisecond accuracy. Timestamped data also makes possible a technically trivial but extremely useful capability: the program can play back a movie of the test, showing exactly how the subject drew the clock, reproducing stroke sequence, precise pen speed at every point, and every pause. This can be helpful diagnostically to the clinician, and can be viewed at any time, even long after the test was taken. Movie playback speed can be varied, permitting slowed motion for examining rapid pen strokes, or sped up for clocks by subjects whose impairments produce vastly slowed motions.


ClockSketch V5.3 - Smith-John-2012-06-03-16-44-11V3.8.csk - 1132%

Examinee Name/Identifier
First: John
Last: Smith
DOB: Jul 23 1945
Eth: nonHisp/Lat Race: White/Cauc
Sex: M Yrs ed.: 12 Handed: R
MM7: 29 MMW: Unk MOC: Unk
Doctor Name: Dana Penney
Test Scored By:
Hospital/Clinic Name: Lahey Clinic
Medical Record ID:
Primary Diagnosis: MCI-amnestic
Primary Diagnosis Confidence: Moderate confidence
Secondary Diagnosis:
Secondary Diagnosis Confidence:
Tertiary Diagnosis:
Tertiary Diagnosis Confidence:
Test Date: Sun 3 Jun 2012 4:44PM EDT
Comments:

○ Copy 0:18
○ Command 0:19
● All

□ Show Points □ Indicate Start Point X □ Extra bold highlight □ Display Classification

☑ Chronological? ☑ Chronological?

4:44:09 PM

Figure 2. The Program Interface.

Movie playback can also be a useful aid when reviewing the classification of strokes in complex clocks.

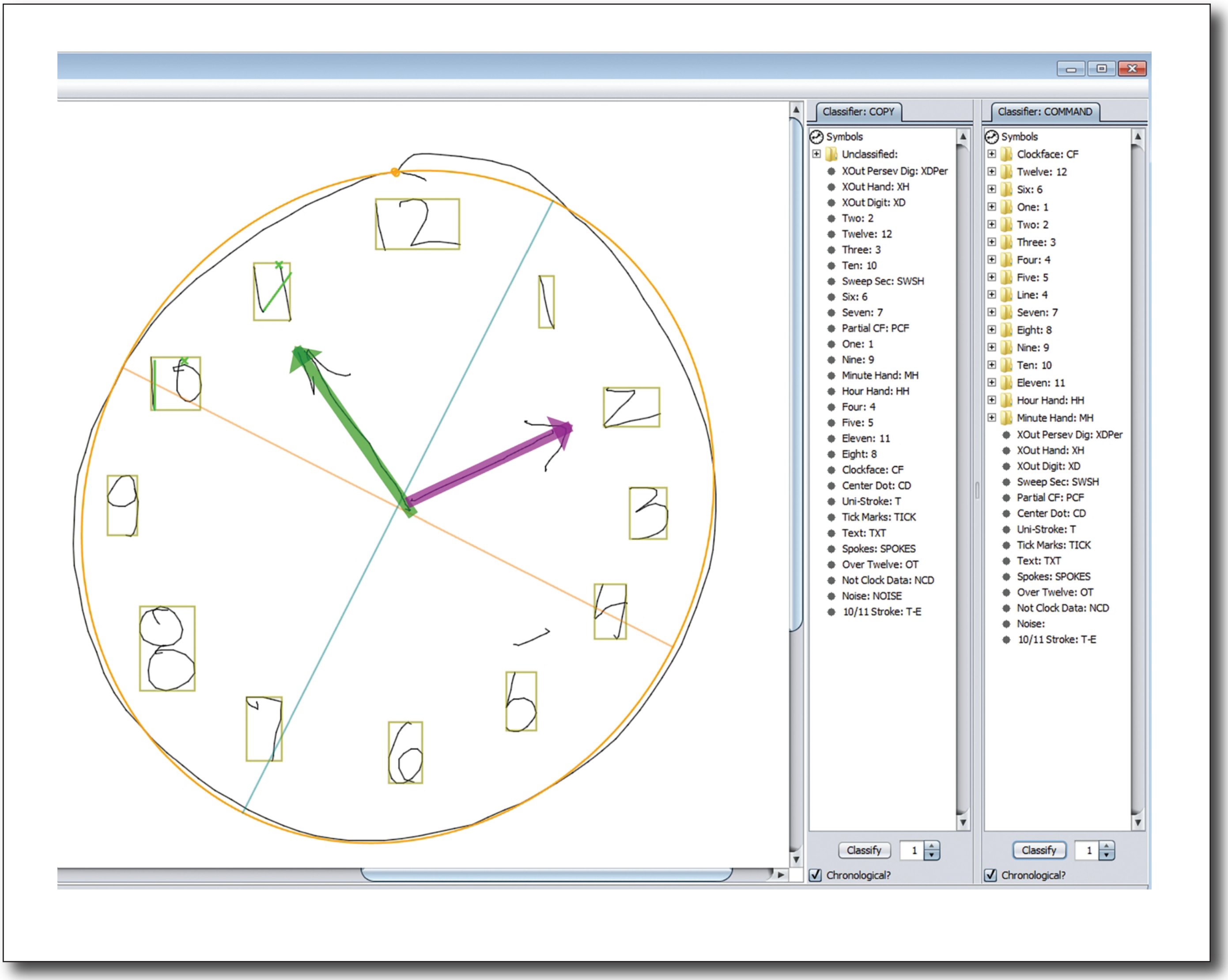
The spatial and temporal precision of the data provides the basis for an unprecedented degree of analysis. The program uses the data to rapidly compute ~500 properties of the drawing that we believe are useful diagnostically (more on this below), including both traditional outcome measures and novel measures of behavior (for example, pauses, pen speed, and others). Because all the measurements are defined in software, they are carried out with no user bias, in real time, at no additional burden to the user.

The program makes it easy to view or analyze its results: it can format the results of its drawing analysis for a single test (that is, the ~500 measurements) as a spreadsheet for easy review; provide comparison metrics on select variables of interest for glanceable clinician review; and can add the data to an electronic medical record. It also automatically creates two versions of each test result: One contains full subject identity data, for the clinician's private records, the second is de-identified and is exported to a central site where we accumulate multiple tests in a standard database format for research.

The program includes in the test file the raw data from the pen, providing the important ability to analyze data from tests given years ago with respect to newly created measures, that is, measurements conceptualized long after that data has been collected. The program also facilitates collection of data from standard psychometric tests (for example, tests of memory, intelligence), providing a customizable and user-friendly interface for entering data from 33 different user-selectable tests.

To facilitate the quality control process integral to many clinical and research settings, the program has a "review mode" that makes the process quick and easy. It automatically loads and zooms in on each clock in turn, and enables the reviewer to check the classifications with a few keystrokes. Clock review typically takes 30 seconds for an experienced reviewer.

The program has been in routine use as both a clinical tool and research vehicle in seven venues (hospitals, clinics, and a research center) around the United States, a group we refer to as the ClockSketch Consortium. The consortium has together administered and classified more than 3500 tests, producing a database of 7000+ clocks (2 per test) with ground-truth labels on every pen stroke.



Classifier: COPY

Symbols

- Unclassified:
  - XOut Persev Dig: XDPer
  - XOut Hand: XH
  - XOut Digit: XD
  - Two: 2
  - Twelve: 12
  - Three: 3
  - Ten: 10
  - Sweep Sec: SWSH
  - Six: 6
  - Seven: 7
  - Partial CF: PCF
  - One: 1
  - Nine: 9
  - Minute Hand: MH
  - Hour Hand: HH
  - Four: 4
  - Five: 5
  - Eleven: 11
  - Eight: 8
  - Clockface: CF
  - Center Dot: CD
  - Uni-Stroke: T
  - Tick Marks: TICK
  - Text: TXT
  - Spokes: SPOKES
  - Over Twelve: OT
  - Not Clock Data: NCD
  - Noise: NOISE
  - 10/11 Stroke: T-E

Classify 1

☑ Chronological?

Classifier: COMMAND

Symbols

- Clockface: CF
- Twelve: 12
- Six: 6
- One: 1
- Two: 2
- Three: 3
- Four: 4
- Five: 5
- Line: 4
- Seven: 7
- Eight: 8
- Nine: 9
- Ten: 10
- Eleven: 11
- Hour Hand: HH
- Minute Hand: MH
  - XOut Persev Dig: XDPer
  - XOut Hand: XH
  - XOut Digit: XD
  - Sweep Sec: SWSH
  - Partial CF: PCF
  - Center Dot: CD
  - Uni-Stroke: T
  - Tick Marks: TICK
  - Text: TXT
  - Spokes: SPOKES
  - Over Twelve: OT
  - Not Clock Data: NCD
  - Noise:
  - 10/11 Stroke: T-E

Classify 1

☑ Chronological?

Figure 3. Partially Classified Clock.

AI Technology Use and Payoff

As noted, our raw input is a set of strokes made up of time-stamped coordinates; classifying the strokes is a sketch interpretation task. Given the range of inputs we must handle (for example, figure 1b, 1c), that interpretation task can be quite challenging.

The program starts by attempting to identify subsets of strokes corresponding to three major elements: the clock circle, digits, and hands. The clock circle is typically the longest circular stroke (or strokes), often but not inevitably drawn first. The program identifies these and fits an ellipse to those points.

Our earliest approach to digit isolation and identification used the clock circle to define an annulus, then segmented the annulus into regions likely to contain each of the 12 numerals. Segmentation was done angularly, by a simple greedy algorithm: the angle to each stroke was measured from the estimated clock center, the angular differences ordered, and the (up to) 12 largest angular differences taken as segmentation points. Strokes in each segment were classified by the angular position of the segment (for example, the segment near the 0-degree position was labeled as a 3).

Hand candidates are identified by finding lines with appropriate properties (for example, having one end near the clock circle center, pointing toward the 11 or 2, and others.).

This extraordinarily simple approach worked surprisingly well for a wide range of normal and slightly impaired clocks.

We have since developed a far more sophisticated approach to digit isolation and identification able to deal with more heavily impaired clocks. It starts by

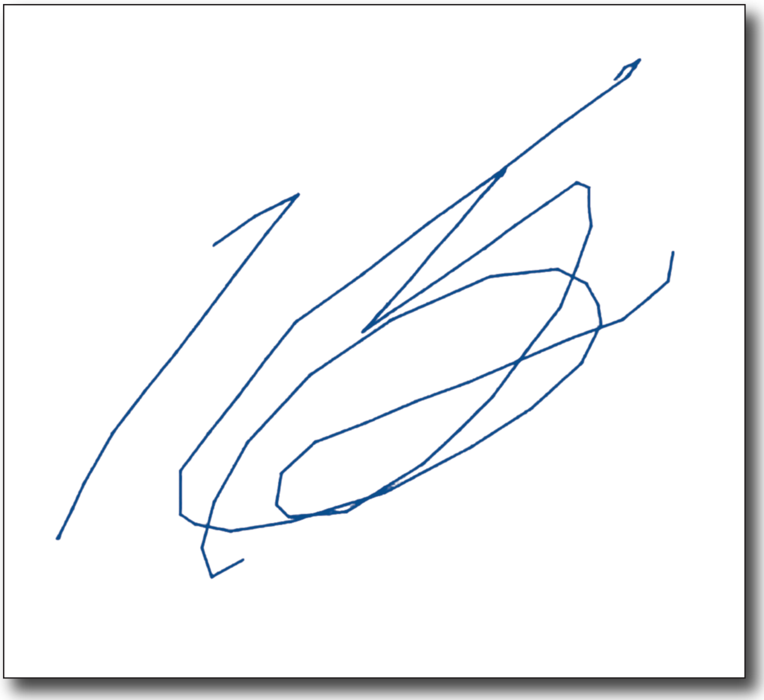


Figure 4. Overwritten Digits.

(12 overwritten with a 6)

using *k*-means to identify strokes likely to be digits, employing a metric combining time (when it was drawn) and distance (stroke midpoint to clock center). The metric is based on the observation that numerals are likely to be further from the clock center and are usually drawn all at once. This set of strokes is divided into subsets likely to be individual digits using a novel representation we call spatiotemporal slices, that combines angle and timing information. In the simpler cases it performs much like the original angle-based segmenter described above, but offers the interesting ability to unpeel layers of ink resulting from crossed out and/or over-written digits, which often produce an incomprehensible collection of ink (for example, figures 1d and 4).

The stroke subsets likely to be individual digits are then identified using a recognizer (Ouyang and Davis 2009) trained on digits from 600 clocks from healthy individuals. The recognizer works from visual features and is thus determining what the strokes look like, independent of how they were drawn (that is, independent of order and timing).

Finally, context matters. Consider the digit in figure 5, from one of our clocks. In isolation it could be interpreted as a 7 or a 1. But when viewed in context (figure 1c), it's interpreted by experienced evaluators as an 8.

We take context into account using a conditional random field that has been trained on triples of angularly sequential digits from both healthy and impaired clocks, enabling the system to classify a digit based in part on the interpretation of the digits on either (angular) side of the current candidate. As just one example, this enables it to classify the digit in figure 5 correctly.

The resulting system has >99 percent digit recognition on clocks from healthy individuals, and 92.3 percent accuracy on impaired clocks of the sort in figuress 1b–1d. For clocks drawn by healthy individuals almost all the error is in segmentation (that is, selecting subsets of strokes belonging to a single numeral), while for impaired clocks segmentation and recognition errors contribute about equally to the 7.7 percent error rate. To date the system also successfully unpacks and interprets about 80 percent of the instances of otherwise incomprehensible layers of ink produced by crossing out or overwriting.

Machine Learning

The clock drawing test is designed to assess diverse cognitive capabilities, including organization and planning, language, and memory, and is employed as a screening test to aid in determining whether performance in any of these areas is sufficiently impaired as to motivate follow-up testing and examination. Our collection of several thousand classified clocks offered an opportunity to use machine learning to determine how effective the hundreds of features computed for each clock might be for making a clinical assessment.

In one early experiment we selected three diagnoses of particular clinical interest and for which we had large enough samples: Alzheimer's ($n = 146$), other dementias ($n = 76$) and Parkinson's ($n = 84$). We used these diagnoses to explore the effectiveness of a large collection of machine-learning algorithms, including SVMs, random forests, boosted decision trees, and others. They were trained to produce binary classifiers that compared each diagnosis against known-healthy subjects (telling us whether the variables would be useful for screening), and each diagnosis against all conditions (telling us whether they would be useful in differential diagnosis).

In general, linear SVM's (table 1) produced the best results. Accuracy rates are good and AUC's are acceptable, but the F1 scores are disappointing in some cases due to low precision rates. As the groups selected for this study have known clinical overlap (for example, Parkinson's and other dementia may have comorbid Alzheimer's), low precision rates may (accurately) reflect this diagnostic overlap. We believe the classifiers may improve as we get additional tests from true positive subjects, and as we learn what additional features may be useful.

One evident follow-up question is, how good is this result? What baseline do we compare it to? It would be best to compare to clinician error rate on judgments made from the same clock tests, but that is not available. There are, however, a number of established procedures designed to produce a numeric score for a clock (for example, from 0–6), indicating where it sits on the impaired versus healthy continuum. We are working to operationalize several of these, which as noted requires making computational something that was written for application by people, and that is often qualitative and vague. Once operationalized, we will be able to use these established procedures to produce their evaluation of all of our clocks. We will then train and test the performance of classifiers using each of those metrics as the test feature, and can then use this as a performance baseline against which to evaluate the results above.

We have also explored the use of a number of data-mining algorithms, including Apriori, FPGrowth, and Bayesian List Machines (BLM) (Letham et al. 2012), in an attempt to build decision models that balance accuracy and comprehensibility. Particularly with BLM, the goal is to produce a decision tree small enough and clear enough to be easily remembered and thus incorporated into a clinician's routine practice.

Clinical Payoff

Our dCDT program has had an impact in both research and practice. The collection and analysis of the wealth of high-precision data provided by the pen has produced insights about novel metrics, particularly those involving time, an advance over traditional clock drawing test evaluation systems, which focus solely on properties of the final drawing (for example, presence/absence of hands, numbers, circle). The new metrics are valuable in daily practice and used by clinicians for their insight into the subject's cognitive status.

Time-dependent variables have, for example, proven to be important in detection of cognitive change. They can reveal when individuals are working harder, even though they are producing normal-appearing outputs. As one example, total time to draw the clock (a measurement not available from the static drawing) differentiates those with amnestic mild cognitive impairment (aMCI) and Alzheimer's disease (AD) from healthy controls (HC) (Penney at al. 2014).

We can also accumulate subtle measures like the total time during the test when the pen is in contact with the paper and is being used to draw (which we call "ink time"), and the converse, the total time during which the subject is not drawing ("think time," composed of the interstroke intervals). Our pilot data suggests that AD can be distinguished from HC by comparing what percent of total test time is spent thinking versus inking, independent of what was drawn (Penney et al. 2013b), suggesting that process can in some cases be more revealing than product.

We have also defined a measure we call pre-first-hand latency (PFHL), measuring the precise amount of time the subject pauses between drawing the numbers on the clock and drawing the first clock hand. This measurement was motivated by the observation that the cognitive demands change during the course of the test: Once started, the process of writing the numerals on the clock is somewhat automatic, but then task demand changes to determining where the hands must be placed. This change in task character inevitably leads to a pause between finishing the numbers and drawing the hands. One interesting early result is that the duration of this pause — the PFHL — appeared to distinguish normal subjects from those with aMCI, AD, and vascular dementia (vAD) (Penney et al. 2011a). We believe PFHL is one measure of decision making, with longer latencies apparent in individuals with cognitive problems (like MCI).

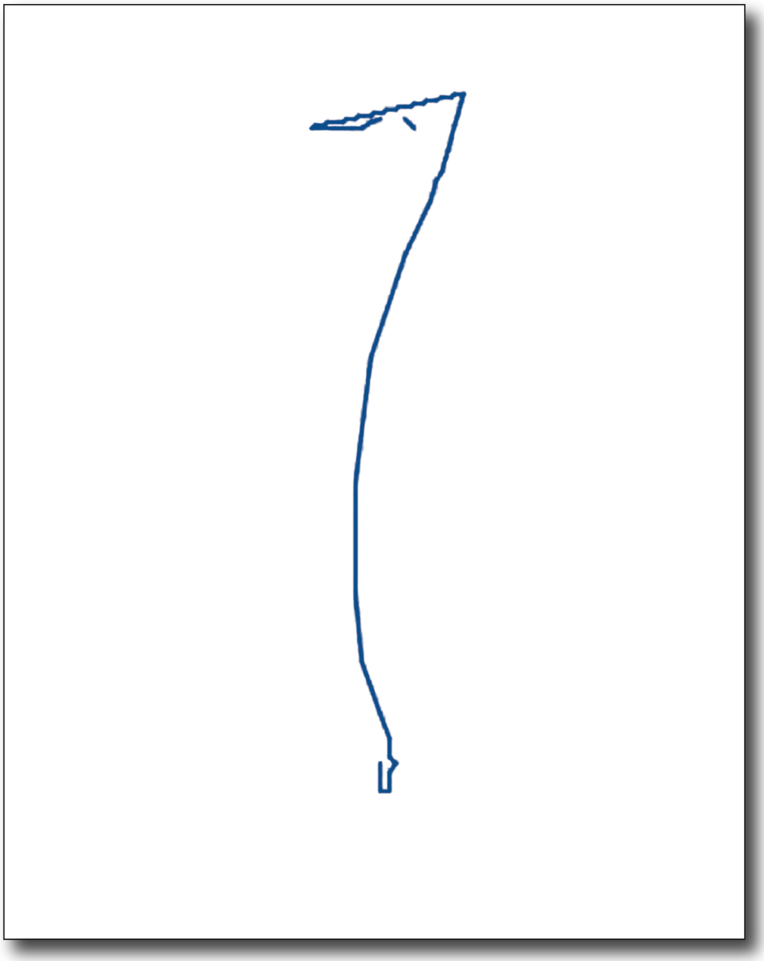


Figure 5. An Ambiguous Digit.

Importantly, our analysis of timing information means we detect these latencies even in drawings that appear completely normal when viewed as a static image.

Latencies are one example of what we have come to call "information between the lines," that is, information arising from what the subject doesn't do, behavior that occurs between one line in the drawing and the next.

Parkinson’s (P) versus Healthy (H)		
	Classified as P	Classified as H
P	56	28
H	52	732
Accuracy	0.908	
F1	0.583	
AUC	0.74	
Dementia (D) versus Healthy		
	Classified as D	Classified as H
D	48	28
H	84	476
Acc	0.824	
F1	0.462	
AUC	0.70	
Alzheimer’s (Az) versus Healthy (H)		
	Classified as Az	Classified as H
Az	132	52
H	64	496
Acc	0.84	
F1	0.69	
AUC	0.76	
Parkinson’s (P) versus All Other (¬P)		
	Classified as P	Classified as ¬P
P	44	40
¬P	204	1740
Acc	0.880	
F1	0.265	
AUC	0.73	
Dementia (D) versus All Other (¬D)		
	Classified as D	Classified as ¬D
D	24	52
¬D	304	1648
Acc	0.824	
F1	0.119	
AUC	0.68	
Alzheimer’s (Az) versus All Other (¬Az)		
	Classified as Az	Classified as ¬Az
Az	104	80
¬Az	168	1676
Acc	0.878	
F1	0.456	
AUC	0.65	

Table 1. SVM Results[2]

Other aspects of diagnostic significance overlooked by traditional feature-driven scoring include the time spent drawing (ink time), total length of stroke (ink length), and overall clock size. In pilot data, subjects with AD worked significantly longer and harder (greater ink time), but produced less output (smaller clocks, smaller ink length) when compared to cognitively intact participants (Penney et al. 2014). This suggests that the level of effort necessary and amount of ink produced are useful for detecting and monitoring cognitive impairment, independent of the drawing accuracy measured by standard clock scoring systems.

Another novel variable we have uncovered concerns the seemingly inadvertently produced ink marks in a clock drawing: these are pen strokes contained within the clock face circle but not clearly identifiable as clock elements. We refer to them as “noise” strokes because they are widely thought to be meaningless and are ignored in traditional analysis. Yet one of our studies (Penney et al. 2011b) found interesting information in the length and location of these strokes: We found that healthy controls made very few of the smallest noise strokes (those <0.3mm), while clinical groups, including those with MCI, made significantly more longer noise strokes, distributed largely in the upper right and left quadrants of the clock (that is, locations on the clock where subjects must negotiate the request to set clock hands to read “10 after 11”). We hypothesize that noise strokes may represent hovering-type marks associated with decision-making difficulty when thinking about where to put the hands to set the time.

All of these features are detected and quantified by the program, producing information that clinicians find useful in practice.

Development and Deployment

For its first seven years this project was produced with quite spare resources (that is, a few small seedling grants). In 2013 we received a year of support from DARPA and have more recently received NSF funding.

The Java code base has (and continues to be) produced by a sequence of talented undergraduate programmers and one computer science professor, inspired, guided, and informed by a few highly experienced neuroscientists who donate their time because they see the opportunity to create a fundamentally new tool for assessing cognitive state.

We started by reconceptualizing the nature of cognitive testing, moving away from the traditional approach of “one test one cognitive domain” (for example, separate tests for memory, executive function, and others.) with standard outcome measures, refocusing instead on the cognitive processes inherent in the drawing task. We broke the complex behavior of clock drawing down to its most basic components (pen strokes), but ensured that we also captured behavioral aspects. We piloted the program at a key clinical site, collecting hundreds of clocks that were used to further refine our set of measurements.

We developed a training program for technicians who administer the test and classify strokes, and adapted the software to both PCs and Macs. We recruited beta testing sites throughout the United States, forming the ClockSketch Consortium. We hosted user training sessions to ensure standard testing and scoring procedures, and measured user proficiency across sites. We established a collaboration with the Framingham Heart Study, a large scale epidemiological study, to enable the development of population-based norms for our measures.[3]

Our ongoing development was significantly aided by a key — and very early design — decision noted above: the raw pen data is always preserved in the test file. This has enabled us over the years to constantly introduce new measurements as we discover more about the precursors to cognitive change. When we do come up with a new measurement, we only have to write the relevant routines to compute it, in effect allowing us to measure the performance of our subjects on factors that were not yet thought of at the time they took the test.

Our deployment strategy has been one of continual refinement, with new versions of the system appearing roughly every six months, in response to our small but vocal user community, which supplied numerous suggestions about missing functionality and improvements in the user interface.

One standard difficulty faced in biomedical applications is approval by internal review boards, who ensure subject safety and quality of care. Here again the use of the digital pen proved to be a good choice: approval at all sites was facilitated by the fact that it functions in the subject's hand as an ordinary pen and presents no additional risk over those encountered in everyday writing tasks.

Next Steps

The digital clock drawing test is the first of what we intend to be a collection of novel, quickly administered neuropsychological tests in the THink project. Our next development is a digital maze test designed to measure graphomotor aspects of executive function, processing speed, spatial reasoning and memory. We believe that use of the digital pen here will provide a substantial body of revealing information, including measures of changes in behavior when approaching decision points (indicating advanced planning), length of pauses at decision points (a measure of decision-making difficulty), changes in these behaviors as a consequence of priming (a measure of memory function), speed in drawing each leg of a solution (measures of learning or memory), and many others.

Capturing these phenomena requires designing mazes with new geometric properties. Because maze completion is a complex task involving the interplay of higher-order cognition (for example, spatial planning, memory), motor operations (pen movement) and visual scanning (eye movement to explore possible paths), little would be gained by simply using a digital pen on a traditional maze. We have designed mazes that will distinguish the phenomena of interest.

We have also created mazes of graded difficulty, accomplished by varying characteristics of the maze, as for example the number of decision points and the presence of embedded choice points. These features will allow us to explore difficulty-tiered decision making by measuring changes in speed approaching a decision point, the length of pauses at each of those points, and by detecting and analyzing errors (for example, back-tracking, repetition of a wrong choice, and others). Tiered decision making is in turn an important measure of executive function that will enable us to detect, measure, and track subtle cognitive difficulty even in correctly solved mazes. We hypothesize that individuals with subtle cognitive impairment, as in MCI and other insidious onset neurologic illness (for example, AD, PD), will pause longer than healthy controls at more difficult decision junctions, while demonstrating only brief or no pauses at easier junctions. We posit that these pauses will be diagnostic even when the correct path solution is chosen. The inclusion of tiered difficulty will allow us to grade cognitive change by assessing cognition at various levels of decision-making difficulty.

The subject will be asked to solve two mazes in sequence, both of which (unknown to the subject) are identical, except that the first has no choice points (added walls remove all choices). This in turn will permit calibrating the effects of priming, giving an indication of the status of memory. The comparison of these two tasks enables using the subject as his/her own control, and using difference scores from the first to second maze will help parse out potential confounding factors (for example, fatigue, depression). These ideas are just the beginning of what appears to be possible with an appropriately designed maze and the data made available with the digital pen.

Larger Implications

One interesting consequence of the detailed data we have is the light it may shed on some previously unknown (or at least under-appreciated) behavioral phenomena that opens up a new approach to understanding cognition. One of these is a phenomenon we call "hooklets." Figure 6 shows a zoomed-in view of an 11, showing that there is a hook at the bottom of the first "1" that heads off in the direction of the beginning of the next stroke. While sometimes visible on paper, hooklets are often less than 0.5mm long, not visible on the paper, yet are clear in the digital record and are detected automatically by our program.

We have hypothesized (Lamar et al. 2011, Penney et al. 2013a) that hooklets represent anticipation: the

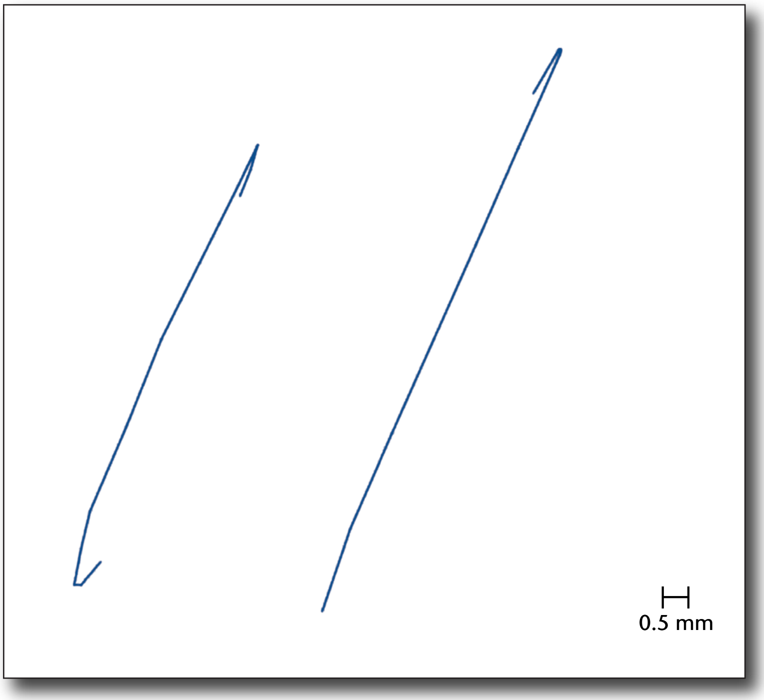


Figure 6. A Hooklet.

subject is thinking about the next stroke and begins moving in that direction before finishing the current one. This is revealing, as the ability to think ahead is a sign of cognitive health: impaired cognition can limit capacity to multitask, leaving resources sufficient only to attend to the current moment. If hooklets are indeed a sign of cognitive health, we have the intriguing possibility that their progressive disappearance may be a (perhaps early) sign of cognitive decline, as in preclinical Alzheimer's.

By focusing on the component processes of cognitive function applied to a standard task and moving away from a traditional approach based on outcome error, we open up the opportunity to study the subtle changes in cognitive health that herald cognitive change before problems manifest. Understanding cognitive strategies that emerge when individuals are consciously or unconsciously compensating for emerging impairment may enable the detection (and hence treatment) of medical conditions far earlier than currently possible, as well as assist with developing new treatments and monitoring their efficacy.

Potential implications of earlier detection and treatment are also receiving increased attention as a result of the increase of life expectancy and resulting graying of populations both in the USA and globally. Current population trends will result in staggering estimated future healthcare costs from even a single form of dementia, Alzheimer's. In the United States in 2012 there were 5.4 million people diagnosed with the disease, costing some $200 billion. The number is projected to rise to 11–16 million by 2050. By then this single disease is projected to result in direct health-care costs in the United States of $1.1 trillion. Given aging populations everywhere, there will be an estimated 135 million afflicted worldwide by 2050, with a corresponding increase in the staggering cost (Alzheimer's Disease International 2013). The additional human costs of care giving, including lost quality of life and suffering, are immeasurable.

The problem is being approached on a number of fronts, including extensive efforts at understanding the biology of the disease and looking for drugs that reverse its effects. But given drug development times (often more than a decade from discovery to approval), methods for early detection that enabled intervention while it was still preclinical or presymptomatic could result in substantial benefits with considerable economic and social value.

Insights About Assessment

Current practice in cognitive assessment typically (and unsurprisingly) assumes that average test scores indicate absence of impairment. We suggest otherwise. We believe that subjects often unwittingly hide early, and thus subtle, impairment behind compensatory strategies, for example thinking harder or working longer in ways that are typically not visible to an observer. Their final results may appear normal (for example, a clock drawing that looks normal), but an ability to see through compensatory strategies would detect the additional mental work and the brief but important additional time spent on a task.

We hypothesize that this can be done by detecting and measuring extremely subtle behaviors produced without conscious effort, as, for example the brief, inadvertent pauses in a task, or the seemingly accidental pen strokes (both noted above), that are normally overlooked or considered spurious and ignored. Detecting and measuring these subtle behaviors reveals the effort normally camouflaged by compensatory strategies (Penney et al. 2014).

We believe this approach to assessment will make possible considerably more detailed information about the cognitive status of an individual, with significant implications for diagnosis and treatment.

Related Work

Over its long history numerous scoring systems have been proposed for the CDT (see, for example, Strauss, Sherman, and Spreen [2006]), but as noted above they may present difficulties by relying on vaguely worded scoring criteria (producing concerns about reliability) or by requiring labor-intensive measurements.

Recent work on automating the clock drawing test is reported by H. Kim (2013), where it was administered on and analyzed on a tablet. That work focused on interface design, seeking to ensure that the system was usable by both subjects and clinicians. It makes some basic use of the timing information available from the tablet, does basic digit recognition and some analysis of subject performance, but is unclear on how much of the subject performance analysis was done by the system versus by the clinician. It does not report dealing with the complexities of the sort noted above, like over-written and crossed out digits, and appears reliant on traditional scoring metrics.

D. Sonntag and colleagues (2013) report on another medical application of the Anoto pen, employing it to annotate medical documents in ways well suited to a pen-based interaction (for example, free-form sketching). The resulting system offers the ease and familiarity of recording information by writing, with the added ability to analyze the annotations and hence integrate them into the medical record.

B. Tiplady and coauthors (2003) reported using the Anoto pen in attempting to calibrate the effects of ethanol on motor control (that is, detecting impaired drivers), by having subjects draw small squares as quickly as possible, a very limited experiment but one that attempted to use data from the pen to detect impaired behavior.

Adapx Inc. developed prototype pen-enabled versions of several standard neuropsychological tests, including trail-making, symbol-digit, and Reys-Osterreith complex figure. Each of these demonstrated the ability to collect digitized data, but did not do data analysis (Salzman, Cohen, and Barthelmess 2010).

Summary

The digital clock drawing test has demonstrated how the original conception and spirit of the clock drawing test can be brought into the digital world, preserving the value and diagnostic information of the original, while simultaneously opening up remarkable new avenues of exploration. We believe the work reported here takes an important step toward a new approach to cognitive assessment, founded on the realization that we no longer have to wait until people look impaired to detect genuine impairment. This offers enormous promise for early differential diagnosis, with clear consequences for both research and treatment.

Acknowledgements

This work has been supported in part by the REW Research and Education Institution, DARPA Contract D13AP00008, NINDS Grant R01-NS17950, NHBLI Contract N01-HC25195, and NIA Grants R01 AG0333040, AG16492, and AG08122. We thank the Anoto Corporation, which from the outset of this project has generously provided technical advice and guidance. The results in table 1 were developed by William Souillard-Mandar.

Notes

[1] Similar technology is available in consumer-oriented packaging from LiveScribe.

[2] The healthy count in the first table differs from that in the second and third in order to ensure age-matched comparisons. Also, as expected, the classes are highly imbalanced. In response we weighted misclassification costs proportionally to class size and found the cost that maximized the AUC (rather than minimized the misclassification rate), using stratified cross-validation.

[3] The Framingham Heart Study began in 1948 with the goal of lifelong physical examinations and lifestyle interviews of the participants every two years to look for patterns related to heart disease. The study's focus has since broadened to other diseases, but the methodology — recruiting and lifelong examination of a large cohort of subjects — means that many of the subjects are healthy, providing an opportunity to establish norms.

References

Alzheimer's Disease International. 2013. The Global Impact of Dementia 2013–2050. London: Alzheimer's Disease International. (www.alz.co.uk/research/GlobalImpactDementia2013.pdf)

Freedman, M.; Leach, L.; Kaplan, E.; Winocur, G.; Shulman, K.; and Delis, D. 1994. *Clock Drawing: A Neuro-Psychological Analysis*. Oxford, UK: Oxford University Press.

Grande, L.; Rudolph, J.; Davis, R.; Penney, D.; Price, C.; Swenson, R.; Libon, D.; and Milberg, W. 2013. Clock Drawing: Standing the Test of Time. In *The Boston Process Approach to Neuropsychological Assessment*. Oxford, UK: Oxford University Press.

Kim H.; 2013. The Clockme System: Computer-Assisted Screening Tool for Dementia. Ph.D. Thesis, Georgia Institute of Technology, College of Computing, Atlanta, GA.

Lamar, M.; Grajewski, M. L.; Penney, D. L.; Davis, R.; Libon, D. J.; and Kumar, A. 2011. The Impact of Vascular Risk and Depression on Executive Planning and Production During Graphomotor Output Across the Lifespan. Abstract presented at 5th Congress of the International Society for Vascular, Cognitive and Behavioural Disorders, 11–14 September, Lille, France.

Letham B.; Rudin C.; Mccormick T.; Madigan D. 2012. Building Interpretable Classifiers with Rules Using Bayesian Analysis. Technical Report Tr609, University of Washington, Department of Statistics.

Nair, A. K.; Gavett, B. E.; Damman, M.; Dekker, W.; Green, R. C.; Mandel, A.; Auerbach, S.; Steinberg, E.; Hubbard, E. J.; Jefferson, A.; and Stern, R. A. 2010. Clock Drawing Test Ratings by Dementia Specialists: Interrater Reliability and Diagnostic Accuracy. *Journal of Neuropsychiatry and Clinical Neurosciences* 22(1): 85–92.

Nasreddine Z. S.; Phillips N. A.; Bédirian V.; Charbonneau S.; Whitehead V.; Collin I.; Cummings J. L.; and Chertkow H. 2005. The Montreal Cognitive Assessment, MOCA: A Brief Screening Tool for Mild Cognitive Impairment. *Journal of the American Geriatrics Society* 53(4): 695–699.

Nyborn, J. A.; Himali, J. J.; Beiser, A. S.; Devine, S. A.; Du, Y.; Kaplan, E.; O'Connor, M. K.; Rinn, W. E.; Denison, H. S.; Seshadri, S.; Wolf, P. A.; and Au, R. 2013. The Framingham Heart Study Clock Drawing Performance: Normative Data from the Offspring Cohort. Clock Scoring. *Experimental Aging Research* 39(1): 80–108.

Ouyang T., and Davis R. 2009. A Visual Approach to Sketched Symbol Recognition. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence,* 1463–1468. Palo Alto, CA: AAAI Press.

Penney, D. L.; Libon, D. J.; Au, R.; Lamar, M.; Price, C. C.; Swenson, R.; Macaulay, C.; Garrett, K. D.; Devine, S.; Delano-Wood, L.; Scala, S.; Flanagan, A.; and Davis, R. 2014. Working Harder but Producing Less: The Digital Clock Drawing Test (DCDT) Differentiates Amnestic Mild Cognitive Impairment and Alzheimer's Disease. Abstract and Poster presented at the 42nd Meeting of the International Neuropsychological Society, Seattle, Washington, 12–15 February.

Penney, D. L.; Libon, D. J.; Lamar, M.; Price, C. C.; Swenson, R.; Scala, S.; Eppig, J.; Nieves, C.; Garrett, K. D.; and Davis, R. 2011a. The Digital Clock Drawing Test (DCDT)-IV: Clock Drawing Time and Hand Placement Latencies in Mild Cognitive Impairment and Dementia. Abstract and poster presented at 5th Congress of the International Society for Vascular, Cognitive and Behavioural Disorders, 11–14 September, Lille, France.

Penney, D. L.; Libon, D. J.; Lamar, M.; Price, C. C.; Swenson, R.; Eppig, J.; Nieves, C.; Garrett, K. D.; and Davis, R. 2011b. The Digital Clock Drawing Test (DCDT) – I: Information Contained Within the Noise. Abstract and poster presented at 5th Congress of the International Society for Vascular, Cognitive and Behavioural Disorders, 11–14 September, Lille, France.

Penney, D. L.; Lamar, M.; Libon, D. J.; Price, C. C.; Swenson, R.; Scala, S.; Eppig, J.; Nieves, C.; Macaulay, C.; Garrett, K. D.; Au, R.; Devine, S.; Delano-Wood, L.; and Davis, R. 2013a. The Digital Clock Drawing Test (DCDT) Hooklets: A Novel Graphomotor Measure of Executive Function. Abstract and Poster presented at the 6th Congress of the International Society for Vascular, Cognitive and Behavioural Disorders, Toronto, Canada 25–28 June.

Penney, D. L.; Libon, D. J.; Lamar, M.; Price, C. C.; Swenson, R.; Scala, S.; Eppig, J.; Nieves, C.; Macaulay, C.; Garrett, K. D.; Au, R.; Devine, S.; Delano-Wood, L.; and Davis, R. 2013b. The Digital Clock Drawing Test (DCDT) in Mild Cognitive Impairment and Dementia: It's a Matter of Time. Abstract and Poster presented at the 6th Congress of the International Society for Vascular, Cognitive and Behavioural Disorders, Toronto, Canada 25–28 June.

Salzman, K.; Cohen, P.; and Barthelmess, P. 2010. Adapx Digital Pen: TBI Cognitive Forms Assessment. Report to US Army Medical Research and Materiel Command, Fort Detrick, MD, July. Washington, DC: US Government Printing Office.

Sonntag D.; Weber M.; Hammon M.; and Cavallo A. 2013. Integrating Digital Pens in Breast Imaging for Instant Knowledge Acquisition. In *Proceedings of the 25th Innovative Applications of Artificial Intelligence Conference,* 1465–1470. Palo Alto, CA: AAAI Press

Strauss E.; Sherman E.; and Spreen O. 2006. *A Compendium of Neuropsychological Tests,* 972–983. Oxford, UK: Oxford University Press.

Tiplady B.; Baird R.; Lutcke H.; Drummond G.; and Wright P. 2013. Use of a Digital Pen to Administer a Psychomotor Test. *Journal of Psychopharmacology* 17(Supplement 3): A71.

Randall Davis is a professor of computer science at the Massachusetts Institute of Technology, where he works on intelligent multimodal interaction and systems for engineering design. He has also been active in the area of intellectual property and software, serving on a number of government studies and acting as an adviser to the court in legal cases. He received his undergraduate degree from Dartmouth College and his Ph.D. from Stanford University. In 1990, he was named a founding fellow of the American Association for Artificial Intelligence and served as president of the association from 1995–1997. From 2012–2014 he served as associate director of MIT's CSAIL. His email address is davis@ai.mit.edu.

David J. Libon is a professor of neurology at Drexel University College of Medicine where he has worked on many problems associated with the assessment of dementia and related neurocognitive disorders. He received his doctoral degree from the University of Rhode Island and went on for training in clinical neuropsychology at the Boston VA Medical Center from 1984–1985. His email address is dlibon@drexelmed.edu

Rhoda Au is a professor of neurology at Boston University School of Medicine and a senior investigator and director of neuropsychology for the Framingham Heart Study. Her research interests are currently focused on cognitive aging and dementia, including early neuropsychological indicators of disease. She received her undergraduate degree from Pomona College and her Ph.D. from the University of California, Riverside. In addition she has an MBA and is working on developing a translational research and technology innovation program with a number of Hong Kong and Mainland China institutions focused on a joint Aging Well Initiative that includes the development of a Chinese national cohort study and entrepreneurial business opportunities. Her email address is rhodaau@bu.edu.

David Pitman is a partner at Kytheram, a consulting firm focused on user experience and interface design, for 10 years. His efforts have spanned mobile, web, and desktop interfaces, as well as human-robot and human-machine interactions. He received bachelor's and master's computer science degrees from the Massachusetts Institute of Technology, focusing on artificial intelligence (undergraduate) and human-computer interaction (graduate). His email address is dpitman@mit.edu

Dana L. Penney is the director of neuropsychology at Lahey Hospital and Medical Center and assistant clinical professor of neurology at Tufts University School of Medicine. She trained in clinical neuropsychology at the West Haven VA Medical Center (1985–1986) and received her doctorate in psychology at the University of Rhode Island in 1988. Her clinical interests are in mild cognitive impairment disorder, the dementias, and the surgical treatment of neurological disorders such as epilepsy and Parkinson's disease. Her clinical work provides the inspiration and foundation for her research, which centers on the detection and measurement of the very earliest possible manifestations of cognitive change and on the development of novel assessment tools.