Mach Learn
DOI 10.1007/s10994-015-5529-5

Learning classification models of cognitive conditions from subtle behaviors in the digital Clock Drawing Test

William Souillard-Mandar[1] · Randall Davis[1] · Cynthia Rudin[1] · Rhoda Au[2] · David J. Libon[3] · Rodney Swenson[4] · Catherine C. Price[5] · Melissa Lamar[6] · Dana L. Penney[7]

Received: 7 February 2015 / Accepted: 5 August 2015

Abstract The Clock Drawing Test—a simple pencil and paper test—has been used for more than 50 years as a screening tool to differentiate normal individuals from those with cognitive impairment, and has proven useful in helping to diagnose cognitive dysfunction associated with neurological disorders such as Alzheimer's disease, Parkinson's disease, and other dementias and conditions. We have been administering the test using a digitizing

Editors: Byron C. Wallace and Jenna Wiens.

✉ William Souillard-Mandar
souillardmandar@csail.mit.edu

Randall Davis
davis@csail.mit.edu

Cynthia Rudin
rudin@mit.edu

Rhoda Au
rhodaau@bu.edu

David J. Libon
dlibon@drexelmed.edu

Rodney Swenson
rod@neuropsychnd.com

Catherine C. Price
cep23@phhp.ufl.edu

Melissa Lamar
mlamar@psych.uic.edu

Dana L. Penney
dana.l.penney@lahey.org

[1] MIT Computer Science And Artificial Intelligence Laboratory, Cambridge, MA, USA

[2] Boston University School of Medicine, Boston, MA, USA

[3] Drexel Neuroscience Institute, Drexel University College of Medicine, Philadelphia, PA, USA

[4] University of North Dakota Medical School, Grand Forks, ND, USA

Published online: 20 October 2015

ballpoint pen that reports its position with considerable spatial and temporal precision, making available far more detailed data about the subject's performance. Using pen stroke data from these drawings categorized by our software, we designed and computed a large collection of features, then explored the tradeoffs in performance and interpretability in classifiers built using a number of different subsets of these features and a variety of different machine learning techniques. We used traditional machine learning methods to build prediction models that achieve high accuracy. We operationalized widely used manual scoring systems so that we could use them as benchmarks for our models. We worked with clinicians to define guidelines for model interpretability, and constructed sparse linear models and rule lists designed to be as easy to use as scoring systems currently used by clinicians, but more accurate. While our models will require additional testing for validation, they offer the possibility of substantial improvement in detecting cognitive impairment earlier than currently possible, a development with considerable potential impact in practice.

Keywords Medical scoring systems · Clock Drawing Test · Cognitive impairment diagnostics · Interpretable machine learning · Machine learning applications

1 Introduction

With progress in medicine extending life expectancy, populations worldwide are "graying," producing a new set of healthcare challenges. As one example, recent estimates suggest that 13.9 % of people above the age of 70 currently have some form of dementia (Plassman et al. 2007), while the Alzheimer's Association projects that by 2050 the number of Americans with Alzheimer's disease will grow to some 13.8 million, with the number worldwide growing to 135 million (Prince et al. 2013). As populations age there will clearly be huge financial, caregiver, and social burdens on our healthcare system and on society in providing care for patients with cognitive impairments.

Research is underway on many fronts, including pharmacological treatment, but there is as yet no cure for cognitive impairments such as Alzheimer's disease (AD) and Parkinson's disease (PD), and pharmaceuticals often take 12 years from discovery to clinical approval. There is however the potential to slow the progress of some forms of cognitive decline, if caught early enough. Hence one important focus of research is early detection.

A variety of tests are used to screen for and assist with differential diagnosis of cognitive decline. One of the simplest and widely used is called the Clock Drawing Test (CDT). In use for over 50 years, it has been a well-accepted cognitive screening tool used in subjects with various dementias and other neurological disorders. The test asks the subject to draw on a blank sheet of paper a clock showing 10 min after 11 (called the Command clock), then asks them to copy a pre-drawn clock showing that time (the Copy clock).

As a simple paper and pencil test, the CDT is quick and easy to administer, non-invasive and inexpensive, yet provides valuable clinical and diagnostic information. It has been shown to be useful as a screening tool to differentiate normal elderly individuals from those with cognitive impairment, and has been effective in helping to diagnose dementias, such as Alzheimer's disease, Parkinson's disease, and other conditions (Freedman et al. 1994; Grande et al. 2013).

[5] University of Florida Gainesville, FL, USA

[6] University of Illinois, Chicago, IL, USA

[7] Lahey Health, Burlington, MA, USA

The CDT is often used by neuropsychologists, neurologists and primary care physicians as part of a general screening for cognitive change (Strub et al. 1985).

But there are drawbacks in the current use of the test. While there are a variety of well-regarded manual scoring systems used by clinicians, these systems often rely on the clinician's subjective judgment of under-specified properties of the drawing. One current scoring system (Nasreddine et al. 2005), for instance, calls for judging whether the clock circle has "only minor distortion," and whether the hour hand is "clearly shorter" than the minute hand, without providing quantitative definitions of those terms, leading to variability in scoring and analysis (Price et al. 2011). Other scoring systems (e.g., Nyborn et al. 2013) specify more precise measures but are far too labor-intensive for routine use.

For the past 7 years neuropsychologists in our group have been administering the CDT using a digitizing pen (the DP-201 from Anoto, Inc.) that, while functioning as an ordinary ballpoint, also records its position on the page with considerable spatial (±0.005 cm) and temporal (12ms) accuracy. The raw data from the pen is analyzed using novel software developed for this task (Davis et al. 2014; Davis and Penney 2014; Cohen et al. 2014); the resulting test is called the digital Clock Drawing Test (dCDT).

The dCDT provides a number of unique capabilities. The spatial precision of the raw data permits the software to do an unprecedented level of geometric analysis of the drawing, with no effort by the user. Because the data points are time-stamped, they capture the entire sequence of behaviors (every stroke, pause or hesitation), rather than just the final result (the drawing). Having time-stamped data means that our software can measure that behavior as well, including informative time intervals like the delay between finishing numbering the clock and starting to draw the hands.

Processing raw data from the pen starts with sketch interpretation, i.e., classifying each pen stroke as one or another component of the clock, e.g., as a minute hand, hour hand, as a specific digit, etc. (Davis et al. 2014).

The next step is clinical interpretation: what does the drawing and the behavior that produced it indicate about the subject's cognitive state? We report here on what light a variety of machine learning techniques shed on answering this question. We describe our work on constructing features that are informative diagnostically, on building classifiers that predict a subject's condition, and on creating classifiers that are both accurate and comprehensible to clinical users.

The medical focus of this paper is on three categories of cognitive impairment chosen because of their clinical significance and because they represent three of the most common diagnoses in our data: memory impairment disorders (MID) consisting of Alzheimer's disease and amnestic mild cognitive impairment (aMCI); vascular cognitive disorders (VCD) consisting of vascular dementia, mixed MCI and vascular cognitive impairment; and Parkinson's disease (PD).

There are two forms of prediction we want to make. *Screening* distinguishes between healthy and one of the three categories of cognitive impairment. For each cognitive impairment category we built models that make a binary-choice prediction indicating whether someone falls in that category or is healthy. We also do a group screening for these three conditions together, i.e., whether a subject falls in any one of the three categories or is healthy. The second task is the diagnosis-like task of clinical group classification—distinguishing one of the three categories from every other of the 43 diagnoses in our data set, including healthy. For brevity in the remainder of the paper we refer to this simply as *diagnosis*.

We define six types of features, detailed in Sect. 3, on which our work is based:

- *Digital-pen features* are the features computed by the dCDT software.
- *Clinician features* are the features used in the existing manual scoring systems created by and used by clinicians;
- *Operationalized clinician features (op-clinician features)* are rigorously defined and computed versions of the clinician features.
- *Simplest features* is a subset of features chosen because we believe they are particularly easy to evaluate by hand, hence less subject to inter-rater variance and usable in the pen-and-paper version of the test.
- *The set of all features* is the union of the digital-pen features, op-clinician features, and simplest features.
- *The MRMR subset of all features* is the first 200 features selected by Minimum-Redundancy–Maximum-Relevance filtering (Peng et al. 2005) from the set of all the features.

We began by using off-the-shelf machine learning methods for their ability to produce accurate predictive models when trained on large amounts of data. Section 4 describes this work and reports on the performance of six classification methods—Gaussian SVM, random forests, CART, C4.5, boosted decision trees, and regularized logistic regression—each of which had access to all features.

These classifiers performed very well in absolute terms, but determining the significance of their performance results requires a baseline to use as a point of comparison. While data are available on the performance of some of the scoring systems used by clinicians (Tuokko et al. 2000; Storey et al. 2001, 2002; Lourenço et al. 2008), these are imperfect measures due to variations in the way the test is given (e.g., whether only one clock is to be drawn, whether the clock face circle is pre-drawn, etc.) and variations in the clinical populations used in evaluation.

To provide a more closely comparable measure of performance, we evaluated our clock test data using seven of the most widely used existing manual scoring systems, selected in a review of the literature. We did this by creating automated versions of these systems, in order to make their use practical for the volume of data we have. One challenge in doing this is that the scoring systems are designed for use by people, and often contain under-specified measures (e.g. deciding whether a clock circle has "only minor distortions.") We thus had to operationalize these algorithms, i.e., specify the computations to be done in enough detail that they could be expressed unambiguously in code. We refer to these as the *operationalized scoring systems*.

One disadvantage of off-the-shelf machine learning classifiers is that they produce black box predictive models that may be impossible to understand as anything other than a numerical calculation. In response, another focus of our work has been on exploring the tradeoff between accuracy and interpretability. In Sect. 6, we provide a definition of interpretability for our problem. We use a recently developed framework, Supersparse Linear Integer Models (SLIM) (Ustun and Rudin 2015; Ustun et al. 2013), and introduce a simple metric to prioritize more understandable features, enabling us to build interpretable linear models.

In Sect. 7, we move to a second class of models consisting of rules and rules lists, built by mining association rules from the data. Some of these rules confirm existing knowledge about correlations between pen-based features and diagnoses, while others appear novel, revealing correlations that had not been reported previously. In a further step in this direction, we constructed rule lists by employing a recently-developed machine learning technique called

Bayesian Rule Lists (BRL) (Letham et al. 2015), which combines associations to create accurate-yet-interpretable predictive models.

Based on the framework outlined above, we carried out a number of experiments that produced the eight primary contributions of this paper:

(i) Starting from a collection of novel clock test features created over the years by members of our team (see e.g., Penney et al. 2011a, b; Lamar et al. 2011; Penney et al. 2013), we created additional single-clock features, as well as features taking advantage of aggregate properties of the clocks and differences between the command and copy clocks. In addition, we operationalized the features used in existing scoring systems, producing the operationalized clinician features, and selected a set of features that we believe to be most easily and reliably determined by clinicians via clinical judgment.

(ii) We show that six state-of-the-art machine learning methods applied to the set of all features produced classifiers with AUC performance ranging from 0.89 to 0.93 for screening and 0.79 to 0.83 for diagnosis. Published AUCs of existing clinician scoring systems (Tuokko et al. 2000; Storey et al. 2001, 2002; Lourenço et al. 2008), which typically attempt only screening (i.e., distinguishing healthy vs. cognitively impaired), range from 0.66 to 0.79 depending on the dataset. Our methods are thus not only significantly more accurate on this task, they are also capable of detecting more fine-grained classes of cognitive impairments for both screening and diagnosis.

(iii) We created operationalizations of seven widely used manual CDT scoring systems, to provide the most direct baseline for evaluating our models. Any free parameters in our operationalized scoring systems were chosen so as to maximize performance of the system, providing an upper bound on the performance of these systems on our data.

(iv) The classifiers produced by the state-of-the-art machine learning methods greatly outperformed the optimized operationalized scoring algorithms for both screening and diagnosis. Where the machine learning methods produced AUCs from 0.89 to 0.93 for screening and 0.79 to 0.83 for diagnosis, the best operationalized scoring algorithms have AUCs of between 0.70 and 0.73 for screening and 0.65 and 0.69 for diagnosis. Thus, using the digital version of the CDT with our machine learning models would lead to more accurate predictions.

(v) We show that applying the machine learning methods to the clinician features leads to models with AUCs from 0.82 to 0.86 for screening and 0.70 to 0.73 for group classification, which is more accurate than the operationalized scoring algorithms. We also show that using the simplest features results in better performance than the operationalized scoring algorithms, with AUCs from 0.82 to 0.83 for screening and 0.72 to 0.73 for group classification. This opens up the possibility of clinicians recording these features and inputting them into our machine learning models, producing more accurate predictions of their patients' conditions, without changing what they attend to in evaluating the test.

(vi) We created Supersparse Linear Integer Models using simplest features, op-clinician features, and the MRMR subset of all features, that are all more accurate than existing scoring systems on the screening task, with AUCs from 0.73 to 0.83 depending on the feature set, and at least as accurate (and often better) on the diagnosis task, with AUCs from 0.66 to 0.77. These models contain very few features and prioritize understandable ones, leading to models that are at least as interpretable as existing algorithms and can be used reliably by clinicians.

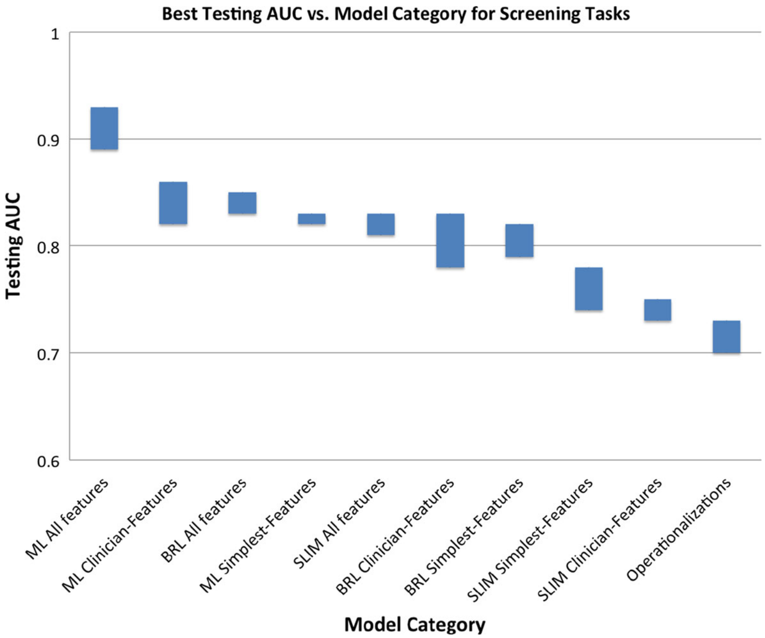

Fig. 1 Summary of results for screening

(vii) We mined association rules and found many that were consistent with existing knowledge about connections between clock features and cognitive conditions, and some that suggested plausible but previously unknown connections.

(viii) We created highly interpretable rule lists using simplest features, op-clinician features, and the MRMR subset of all features, resulting in classifiers with AUCs ranging from 0.78 to 0.85 for screening and 0.69 to 0.74 for diagnosis, depending on the feature set and condition. As above, these models might be usable by clinicians at least as easily, and possibly more reliably and accurately, than existing scoring systems.

Figures 1 and 2 summarize the results described above, showing the range of accuracies achieved by our different models for screening and diagnosis, respectively, ordered by decreasing upper bound. Each model category is a pairing of a class of model (traditional machine learning models, Supersparse Linear Integer Models, or Bayesian Rule Lists) with a feature set (simplest features, clinician features, or all features/MRMR subset of all features). Each bar shows the range of the AUC's across test folds for each condition, for the best algorithm in each category. For example, on the screening plot, "ML All features" indicates the range of accuracies of the best machine learning algorithms using all features, over the four possible screening tasks.

1.1 Related work

The goal of creating interpretable models—described in early expert systems work as the need for transparency—has received considerably less attention in the last two decades. In that time, machine learning methods have rarely been used for creating scoring systems; instead these systems are often created manually by teams of domain experts, or created by heuristically rounding logistic regression coefficients. There are some recent efforts to

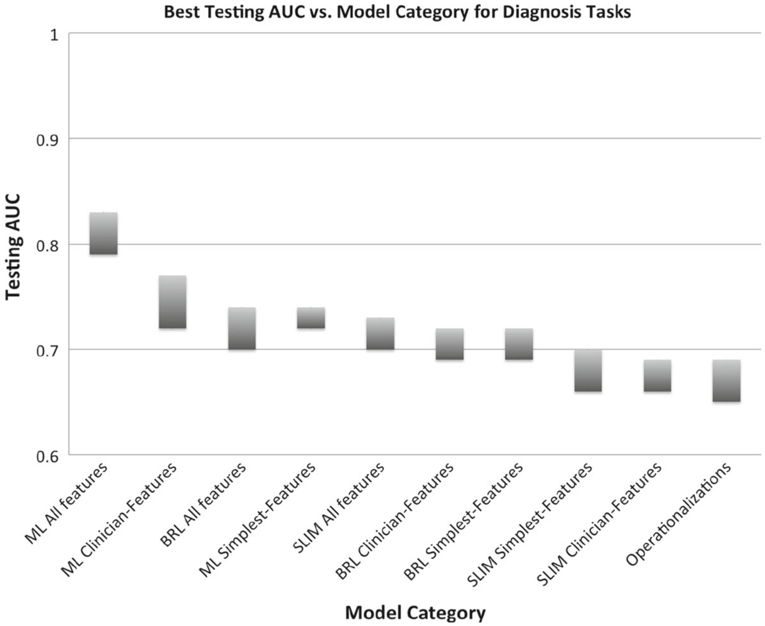

Fig. 2 Summary of results for diagnosis

develop and use machine learning in domains such as credit scoring (Martens et al. 2007), crime prediction (Steinhart 2006; Andrade 2009; Ridgeway 2013), national defense (ABS Consulting 2002), marketing (Hauser et al. 2010; Verbeke et al. 2011; Wang et al. 2015), medical diagnosis (Tian and Tibshirani 2011; Van Belle et al. 2012; Letham et al. 2015; Wang and Rudin 2015), and scientific discovery (Sun 2006; Freitas et al. 2010; Haury et al. 2011). None of these works use a machine learning algorithm designed for producing scoring systems, like the SLIM method we use in this paper. SLIM creates linear models that can be heavily customized through integer programming techniques. This means that it can handle constraints on the coefficients, constraints on the false positive rate, or very customized definitions of interpretability that no other current method can handle. Bayesian Rule Lists, on the other hand, creates logical IF-THEN rule models. Bayesian Rule Lists is a competitor for CART (Classification and Regression Trees, Breiman et al. 1984). CART uses a greedy (myopic) optimization method to produce decision trees, whereas BRL makes a statistical approximation to reduce the size of the problem, and then aims to fully solve the reduced problem for the best rule list. This means that for moderately sized problems like ours, BRL will generally produce solutions that are more accurate and sparse than CART.

Numerous papers in the clinical literature describe a variety of manual scoring systems for the clock test (Manos and Wu 1994; Royall et al. 1998; Shulman et al. 1993; Rouleau et al. 1992; Mendez et al. 1992; Borson et al. 2000; Libon et al. 1993; Sunderland et al. 1989), none of which used a machine learning approach to optimize for accuracy.

There have also been a few attempts to create novel versions of the clock drawing test. The closest work to ours (Kim et al. 2011a, b) builds a tablet-based clock drawing test that allows the collection of data along with some statistics about user behavior. However, that work focuses primarily on the user-interface aspects of the application, trying to ensure that it is usable by both subjects and clinicians, but not on automatically detecting cognitive conditions.

No work that we know of—and certainly none used in practice—has used state-of-the-art machine learning methods to create these systems or has reported levels of accuracy comparable to those obtained in this work. In addition, no work that we know of has aimed to understand the tradeoff between accuracy of prediction and interpretability for the clock drawing test.

2 The digital Clock Test data

Over the past 7 years we have accumulated a database of 3541 digital clock tests whose strokes have been carefully classified and independently reviewed for accuracy. Some subjects have been tested multiple times over the years; to avoid issues that might arise from repeated exposure to the test, only the first test for each subject was included in our analysis, resulting in 2169 tests (each of which has both a Command and Copy clock, yielding 4338 distinct drawings).

Our dataset consists of subjects with diverse medical conditions. The focus in this paper is on three categories of cognitive impairment chosen because of their clinical significance and because they represent three of the most common diagnoses in our data, along with healthy controls:

- The memory impairment disorders (MID) group consists of 206 subjects diagnosed as having Alzheimer's disease or amnestic MCI on the basis of one or more criteria: consensus diagnosis of a neuropsychologist and neurologist, neuropsychological test finding, or selection into a study following the research diagnostic criteria. Alzheimer's disease is the most common form of dementia, accounting for 60–70 % of dementia cases (Petersen et al. 2014). Mild cognitive impairment (MCI) can present with a variety of symptoms; when memory loss is the predominant symptom it is termed "amnestic MCI" and is frequently seen as a transitional stage between normal aging and Alzheimer's disease (Albert et al. 2011). We would expect memory problems on the clock test but do not expect significant motor slowing during the early stages of the disease. In our sample, subjects with amnestic MCI meet criteria established by Petersen et al. (2014) and have circumscribed memory loss in the context of otherwise intact cognition and no report of functional problems. Our subjects with Alzheimer's disease are primarily at an early stage of the disease.
- The vascular cognitive disorders (VCD) group consists of 121 subjects diagnosed with vascular dementia, mixed MCI, or vascular cognitive impairment (VCI). Vascular dementia is widely considered the second most common cause of dementia after Alzheimer's disease, accounting for 10 % of cases (Battistin and Cagnin 2010). Early detection and accurate diagnosis are important, as risk factors for vascular dementia are important targets for medical intervention. We hypothesize motor and cognitive slowing effects on the test performance.
- The PD group has 126 subjects diagnosed with Parkinson's disease. Early in the course of the disease the most obvious symptoms are movement-related and may include tremor, rigidity, slowness of movement and difficulty with gait. Later, thinking and behavioral problems may arise, with dementia (if diagnosed) most often occurring in the advanced stages of the disease. There is no cure yet, but medical and surgical treatments are effective at managing the motor symptoms of the disease.

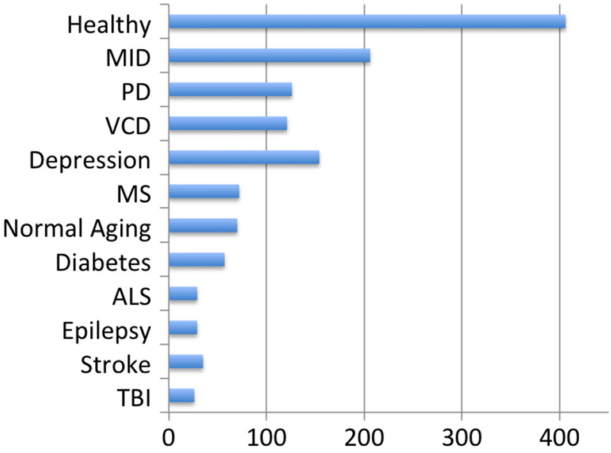

Fig. 3 Histogram of most frequent conditions in the data set

- Our set of 406 healthy controls (HC) comes from people who have been longitudinally studied as participants in the Framingham Heart Study[1] and are judged by trained examiners to be cognitively intact.

The remainder of the tests have other neurological, psychiatric, and medical conditions; the distribution of the most frequent conditions is shown in Fig. 3.

Figure 4 illustrates representative clock drawings from our dataset from a subject from the HC group, a subject in the memory impairment group, and a subject diagnosed with PD. As the figure suggests, clocks by HC subjects are typically reasonably round, have all 12 digits present and spaced regularly around the clock, and have hands pointing towards digit 11 and digit 2. Hands often have arrowheads, and the minute hand is often but not invariably longer than the hour hand, following the traditional clock format. A center dot is also common.

There are many possible variations found in both HC and impaired subjects.

- Clocks vary significantly in size, with some subjects drawing them much smaller (Fig. 4c).
- There may be a gap between the start and the stop of the clockface (Fig. 4c).
- Digits maybe be missing, crossed-out, repeated, or with poor angular spacing (Fig. 4b).
- Digits greater than 12 are sometimes drawn.
- Hands can be missing (Fig. 4b), crossed-out, or repeated, with arrowheads sometime pointing toward the clock center.
- Some clocks contain stokes used by subjects for spatial arrangement, and tickmarks used as replacement for digits.
- Subjects sometime use additional text in their drawings, for example to write the time as a memory aid or in lieu of a number.
- We have defined "noise" as strokes that are not part of the representation of defined clock elements (e.g. hand, digit) but are clearly produced during the drawing process and are

[1] Initiated in 1948, the Framingham Heart Study (FHS) originated with biennial examinations of an original cohort to identify the determinants of heart disease and stroke. The focus has since broadened to include additional generations of participants and expanded investigations to other types of chronic diseases. The majority of their participants remain cognitively intact. Beginning in 2011, FHS adopted the dCDT as part of its cognitive test suite.

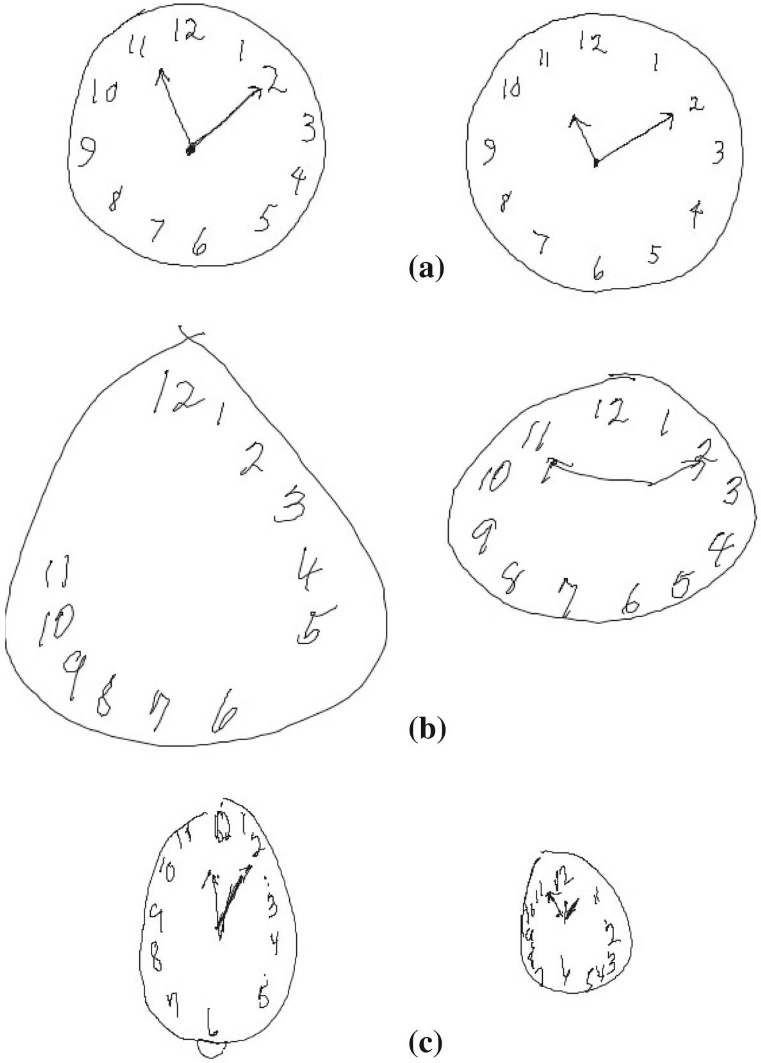

Fig. 4 Example clocks, to scale, from our dataset for healthy controls, Alzheimer's disease, and Parkinson's disease, with command clock on the left and copy clock on the right. **a** Healthy Control, **b** Alzheimer's disease, **c** Parkinson's disease

intentional (i.e. not random pen drops) (Penney et al. 2011a). They vary from tiny dots to longer lines (Fig. 4c).

- A more subtle feature, hooklets (Lamar et al. 2011; Penney et al. 2013), can also be observed. These are abrupt changes in the direction at the end of a stroke that head toward the beginning of the next stroke. For example, when drawing the numbers on a clock, subjects may leave a "tail" on the end of one digit stroke that points toward the start of the first stroke of the next digit.

As described elsewhere (Davis et al. 2014), our software analyzes the raw data from the pen, automatically classifying strokes as part of the clock face circle, hands, numbers, etc.

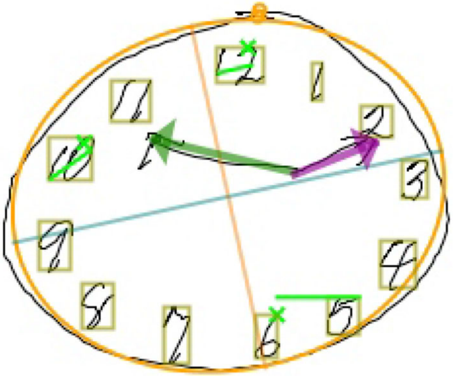

Fig. 5 Example classified command clock from Fig. 4b. An *ellipse* is fit to the clockface, with the major and minor axis shown; *bounding boxes* are drawn around each digit; *arrows* show the overall direction of the hands; the *lines on digits 5*, *10*, and *12* show hooklets, with “x”s indicating the start of the next stroke after each hooklet. The system adds the colored overlays as a way of making stroke classification visually obvious

It also permits assistance from the user, needed in difficult cases (e.g., for clocks by more impaired subjects). Figure 5 shows a screenshot of the system after the strokes in a clock have been classified, showing the starting point for the work reported here.

Stroke classification is a key first step, as it enables appropriate measurement of clock features, e.g., the average size of the numerals, how accurately the hands show the time, the latency between finishing drawing the numerals and starting to draw the hands, etc. (see, e.g., Davis et al. 2011). The spatial and temporal accuracy of the pen data permits our system to make precise measurements that are implausibly difficult with ordinary ink on paper.

3 Feature construction

We constructed five sets of features to use with the various algorithms we employed. We describe the feature sets and their objectives below.

3.1 Digital-pen features

These are the features computed by the dCDT software along with additional features based on those. They fall into four categories.

3.1.1 Single-clock-measurements

These are measurements of geometric or temporal properties of components of a single clock. For example:

- The number of strokes, the total ink length, the time it took to draw, and the pen speed for each component (e.g. the clockface, all digits, and all hands).
- The length of the major and minor axis of the fitted ellipse as well as the distance and angular difference between starting and ending points of the clock face (Fig. 6A).
- Digits that are missing or repeated, the height and width of their bounding boxes (Fig. 6B).
- Omissions or repetitions of hands, the hour hand to minute hand size ratio, the presence and direction of arrowheads, and angular error from their correct angle (Fig. 6C).
- Whether the minute hand points to digit 10 instead of digit 2, which can happen as a consequence of the instruction to set the time to “10 past 11”.

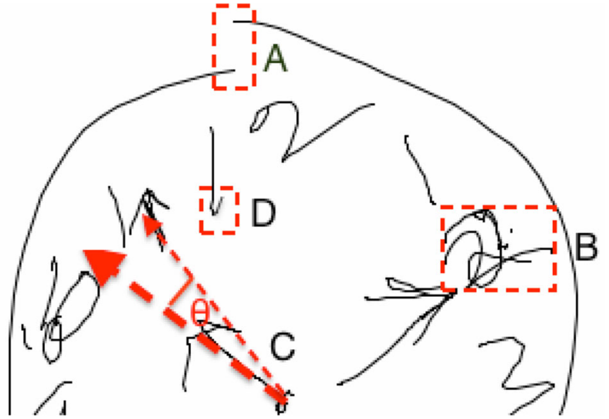

Fig. 6 *A* the distance between starting and ending point of the clockface, as well as the angular difference; *B* digit repetition; width and height of the bounding box; *C* the difference in angle between a hand and its correct angle; *D* hooklet presence, length, and direction

- The presence, length, and direction of hooklets (Fig. 6D).
- The presence of tick marks, spokes, any text, or a center dot for the hands.
- The number of and length of noise strokes.
- Timing information is used to measure how quickly different parts of the clock were drawn. One particularly interesting latency feature is one called the pre-firsthand latency, the time that elapsed between the first stroke of the first hand drawn and whatever was drawn previously (e.g. Penney et al. 2011b).

3.1.2 Single-clock-aggregates

These are aggregates of geometric or temporal properties of a single clock. For example:

- The total time to draw the clock and the total number of strokes used.
- The average height, average width, and average length of all digits present.
- The number of digits missing or repeated.
- Whether digits 12, 6, 3, and 9 are drawn before any other digits.[2]
- Measures of the distribution of digits around the clock. For example, one feature counts the number of digits in the clock that do not have 2 other digits within 45° on either side; another feature reports whether all non-anchor digits are in the correct octant of the clock circle; yet another reports the variance in the distance of digits from the clockface.
- The percentage of time spent drawing versus thinking (holding the pen off the paper) for one clock.

3.1.3 Both-clock-aggregates

These are aggregates over both the command and the copy clock. For example:

- The total time to draw both clocks.

[2] When the 12, 3, 6, and 9 digits are drawn before any other digits, they are referred to as "anchor digits," as they are being used to organize the drawing.

- The total number of strokes used.
- The average height, average width, and average length of all digits present in both clocks.
- The number of digits missing in both clocks.
- The percentage of time spent drawing versus thinking for both clocks.

3.1.4 Clock differences

We computed the difference in value of a feature across the command clock and the copy clock, e.g, the difference in the total time to draw each clock. This follows the intuition that because the command and copy clocks engage different cognitive functions, differences between them may be revealing.

3.2 Clinician features and operationalized-clinician features (op-clinician features)

Some of the features found in manual scoring systems (the "clinician features") are quantitative, such as checking for the presence of a digit or a hand. Others are less well defined: for example, one feature calls for determining whether the minute hand is "obviously longer" than the hour hand, while another checks whether there are "slight errors in the placement of the hands." These can be estimated by a clinician, but it is not immediately obvious how to compute them in a program in a way that captures the original intent. Section 5 describes our efforts to create the operationalized versions of these features.

The operationalized features then allow us to create computable versions of the manual scoring systems, providing a baseline against which to compare the classifiers we build. We also use these features with the machine learning algorithms to measure how predictive they can be in models of other forms.

3.3 Simplest features

This is a subset of the features available in the traditional pen-and-paper version of the test, selecting those for which we believe there would be little variance in their measurement across clinicians. We expect, for example, that there would be wide agreement on whether a number is present, whether hands have arrowheads on them, whether there are easily noticeable noise strokes, etc.

Models created using this set of features would be applicable to the traditional pen-and-paper version of the test (i.e. without the digitizing pen), with clinicians likely to be able to measure the features more consistently than those in existing scoring systems.

3.4 All features

This is the union of digital-pen features, op-clinician features, and simplest features. Our intent here is to build the best model possible, without regard to the number of features, their interpretability, etc., in order to get the maximum benefit from the data.

3.5 MRMR subset of all features

From among all of the features, we created a subset of the first 200 selected by Minimum-Redundancy–Maximum-Relevance filtering (Peng et al. 2005). This set of features can be used when it would be computationally too expensive to use the set of all features.

Table 1 Classification results for the screening task: distinguishing clinical group from HC

Algorithm	Memory impairment disorders versus HC	Vascular cognitive disorders versus HC	PD versus HC	All three versus HC
C4.5	0.75 (0.08)	0.72 (0.07)	0.75 (0.06)	0.78 (0.08)
CART	0.78 (0.07)	0.75 (0.13)	0.76 (0.10)	0.76 (0.10)
SVM Gaussian	**0.89 (0.06)**	**0.84 (0.08)**	**0.86 (0.08)**	**0.91 (0.09)**
Random forest	**0.89 (0.10)**	**0.88 (0.09)**	**0.91 (0.11)**	**0.89 (0.06)**
Boosted decision trees	**0.93 (0.09)**	**0.88 (0.11)**	**0.87 (0.08)**	**0.90 (0.12)**
Regularized logistic regression	**0.88 (0.11)**	**0.85 (0.07)**	**0.91 (0.08)**	**0.89 (0.09)**

Each entry in the table shows the mean and standard deviation AUC of a machine learning algorithm across 5 folds. The first column is for the task of distinguishing memory impairment disorders versus HC, the second column is for vascular cognitive disorders versus HC, the third column is for PD versus HC, and the last column is for any of the three cognitive impairments versus HC. F-scores are in "Appendix 2", Table 18

4 Machine learning on all features

Our aim in this section is to determine the highest accuracy attainable for classifiers built from our data, by applying state-of-the-art machine learning methods to the set of all features.

We began with the screening task, seeking to develop classifiers able to distinguish HC subjects from those with one of the conditions listed earlier: memory impairment disorders, vascular cognitive disorders, and PD, as well as whether the subject is a HC or falls under any of the three clinical diagnosis groups.

We generated classifiers using multiple machine learning methods, including CART (Breiman et al. 1984), C4.5 (Quinlan 1993), SVM with gaussian kernels (Joachims 1998), random forests (Breiman 2001), boosted decision trees (Friedman 2001), and regularized logistic regression (Fan et al. 2008). We used stratified cross-validation to divide the data into 5 folds to obtain training and testing sets. We further cross-validated each training set into 5 folds to optimize the parameters of the algorithm using grid search over a set of ranges. "Appendix 1" provides additional details the implementations.

Table 1 shows the prediction quality for all of the machine learning algorithms we used, reported as the mean and standard deviation of performance over the test folds. We chose to measure quality using area under the receiver operator characteristic curve (AUC) as a single, concise statistic; we display full ROC curves in Fig. 7. Each curve is a mean over the 5 folds, with 95 % confidence intervals displayed as bars along the curves. We assessed statistical significance for the experiments in Table 1 using matched pairs t-tests; bold indicates algorithms whose result was not statistically significantly different from the best algorithm.[3] In addition, we provide F-scores for all of our experiments in "Appendix 2". Note that no single machine learning method can be declared the winner across all experiments.

The best classifiers achieve AUC measures from the high 80s to the low 90s. With this level of prediction quality, these methods can be quite helpful as decision aids for clinicians.

For our sample of subjects, these results are superior to published accuracies of existing scoring systems, even where those scoring systems focused on the simpler screening task of

[3] These hypothesis tests are problematic because experiments between folds are not independent, but there is apparently no good alternative for testing (see, for instance, Markatou et al. 2005).

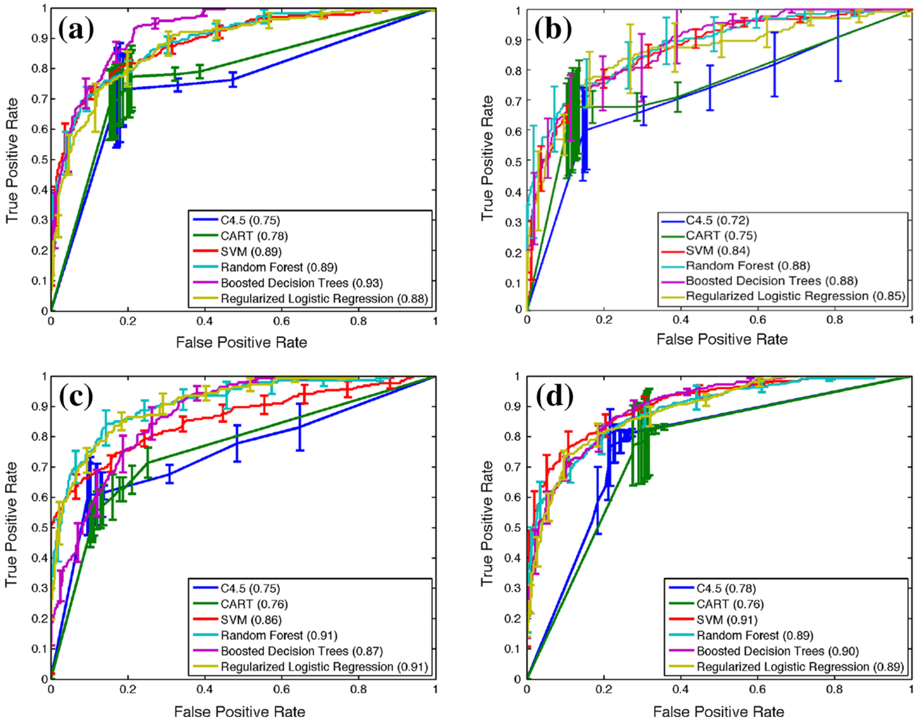

Fig. 7 ROC curves for screening task (Table 1). **a** Memory impairment disorders versus HC, **b** vascular cognitive disorders versus HC, **c** PD versus HC, **d** all three versus HC

distinguishing HC subjects from those with any form of cognitive impairment, instead of the more fine-grained categories we use. Published results report AUC levels ranging from 0.60 to 0.79 (Tuokko et al. 2000; Storey et al. 2001, 2002; Lourenço et al. 2008), with variance in the performance across reports. As an example of the ranges and variance, AUC accuracy for two widely used scoring systems have been reported from 0.66 (Storey et al. 2002) to 0.79 (Storey et al. 2001) for Shulman (Shulman et al. 1993), and from 0.7 (Storey et al. 2002) to 0.78 (Storey et al. 2001) for Mendez (Mendez et al. 1992).

To produce the full ROC curves shown in Fig. 7 for a particular model (machine learning model, or scoring system), we rank subjects according to their score in the model and build the curve from the left (subjects with the highest score) to right (subjects with the lowest score). This way, the left part of the curve represents subjects most likely to have an impairment.

The second set of experiments was aimed at diagnosis, i.e., distinguishing subjects in one of our clinical groups from subjects who have any other medical, neurological, or psychological condition. Table 2 shows comparative accuracy results; Fig. 8 shows the associated ROC curves. As expected, diagnosis is a more difficult task, leading to the best algorithms having AUC's within the high 70s to low 80s.

5 Operationalized scoring systems

Having established performance for machine learning classifiers, we would like to know how they compare to the models currently in use by clinicians. Ideally, we would determine

Table 2 Classification results for the diagnosis task: distinguishing one cognitive impairment group from all other diagnoses

Algorithm	Memory impairment disorders versus all others	Vascular cognitive disorders versus all others	PD versus all others	All three versus all others
C4.5	0.71 (0.10)	0.67 (0.06)	0.71 (0.09)	0.66 (0.09)
CART	0.72 (0.06)	0.69 (0.09)	0.68 (0.09)	0.64 (0.04)
SVM Gaussian	**0.79** (**0.07**)	**0.77** (**0.13**)	**0.81** (**0.11**)	0.72 (0.06)
Random forest	**0.83** (**0.06**)	**0.79** (**0.10**)	**0.81** (**0.07**)	0.73 (0.04)
Boosted decision trees	**0.80** (**0.09**)	**0.77** (**0.08**)	**0.77** (**0.09**)	**0.82** (**0.05**)
Regularized logistic regression	**0.78** (**0.06**)	**0.79** (**0.05**)	**0.82** (**0.05**)	**0.79** (**0.07**)

Each entry in the table shows the AUC and standard deviation of a machine learning algorithm for distinguishing one disease from the others. For instance, the entry in the table corresponding to memory impairment disorders and C4.5 indicates the accuracy in distinguishing memory impairment disorders from every other condition. F-scores are in "Appendix 2", Table 19

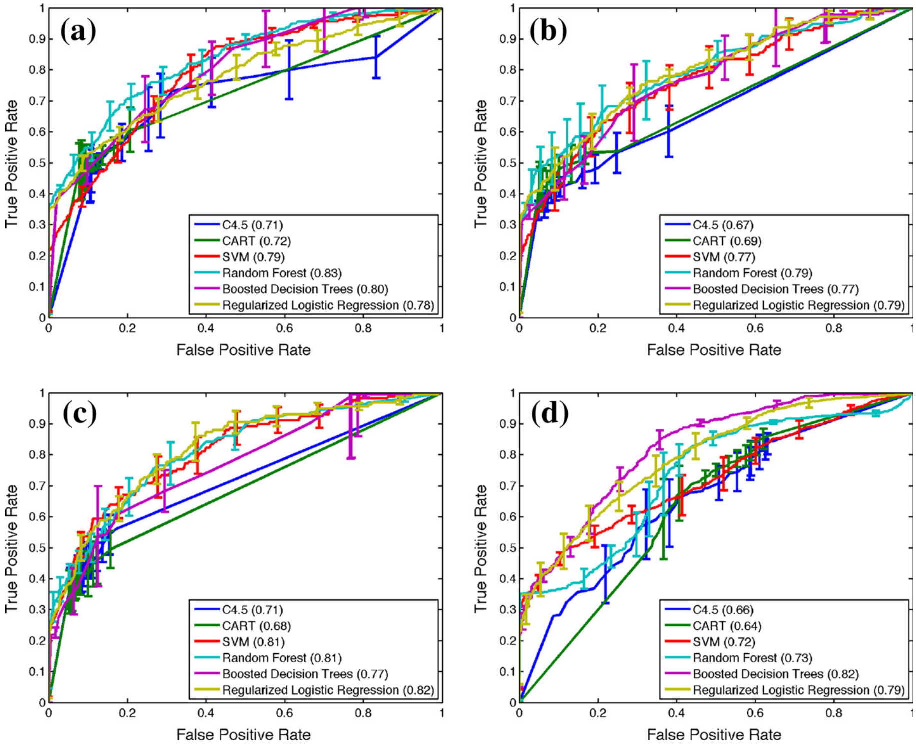

Fig. 8 ROC curves for the diagnosis task (Table 2). **a** Memory impairment disorders versus all others, **b** vascular cognitive disorders versus all others, **c** PD versus all others, **d** all three versus all others

this by having a large number of our clock tests manually evaluated by clinicians using the scoring systems in current use, but this was not pragmatically possible. We were, however, able to establish a useful baseline by creating computational models of the existing scoring

Table 3 Original Rouleau scoring system (Rouleau et al. 1992)

Maximum: 10 points
1. Integrity of the clockface (maximum: 2 points)
2: Present without gross distortion
1: Incomplete or some distortion
0: Absent or totally inappropriate
2. Presence and sequencing of the numbers (maximum: 4 points)
4: All present in the right order and at most minimal error in the spatial arrangement
3: All present but errors in spatial arrangement
2: Numbers missing or added but no gross distortions of the remaining numbers
Numbers placed in counterclockwise direction
Numbers all present but gross distortion in spatial layout
1: Missing or added numbers and gross spatial distortions
0: Absence or poor representation of numbers
3. Presence and placement of the hands (maximum: 4 points)
4: Hands are in correct position and the size difference is respected
3: Sight errors in the placement of the hands or no representation of size difference between the hands
2: Major errors in the placement of the hands (significantly out of course including 10–11)
1: Only one hand or poor representation of two hands
0: No hands or perseveration on hands

systems, resulting in models which we call *operationalized scoring systems*. The goal here was to create automated versions of the scoring systems used by clinicians so that we could reproduce the judgments they would make when applying one of the existing scoring systems.

There are a variety of scoring systems for the clock test, varying in complexity and the types of features they use. In each of the systems, points are added and subtracted based on features of the clock, such as whether clock hands are present, digits are missing, or the correct time is shown. A threshold is then used to decide whether the test gives evidence of impairment.

We worked with clinicians to identify the most widely used scoring algorithms, leaving us with seven: Manos (Manos and Wu 1994), Royall (Royall et al. 1998), Shulman (Shulman et al. 1993), Rouleau (Rouleau et al. 1992), Mendez (Mendez et al. 1992), MiniCog (Borson et al. 2000), and Libon (Libon et al. 1993) (based on Sunderland et al. 1989). Table 3 shows the Rouleau scoring criterion; we focus on it as an example of the operationalization process.

To operationalize these systems, we had to transform relatively vague terms, such as "slight errors in the placement of the hands" and "clockface present without gross distortion", into precise rules that can be programmed. One challenge was defining what was meant by the vague terms.

As one example of our approach, guided by the clinicians, we translated "slight errors in the placement of the hands" to "exactly two hands present AND at most one hand with a pointing error of between ϵ_1 and ϵ_2 degrees", where the ϵ_i are thresholds. Similarly, "clock

Table 4 Operationalized non-obvious features for Rouleau

Variable	Description
Eccentricity of fitted ellipse	$\sqrt{(1-(\frac{b}{a})^2)}$ where a and b are half the major and minor axes respectively. A perfect circle has value 0, the value increases toward 1 as it gets flatter
ClockfaceClosedPercentage	The percentage of the angle of the clockface that is closed
DigitsAngleError	The average angle error of digits from their correct angle. A measure of the distribution of digits angularly
DigitNeighborsTest	A count of the number of digits in the clock with fewer than 2 other digits within ±45°. A second measure of the distribution of the digits angularly
HandAngleError	The difference in angle between the hand and the digit it should point to
HandRatio	The ratio: length of the hour hand/length of minute hand

face present without gross distortion" became "eccentricity of the clockface $\leq \epsilon_3$ AND clock face closed percentage $\geq \epsilon_4$".

Table 4 shows the non-obvious features used in the Rouleau scoring system (e.g. "digit missing" is obvious), while Table 5 shows the resulting operationalized scoring system. Operationalized scoring systems for all the other manual scoring systems are given in "Appendix 3".

The clinicians on our team confirmed the form and content of these operationalized scoring systems and provided initial values for the thresholds which they believed made the operationalizations capture the intent of the original manual scoring systems. For instance, the initial hand pointing thresholds were 15° and 30°.

Starting from these initial values, we created a range of possible values for each parameter (see "Appendix 4"), then selected parameter values via a 5-fold stratified cross-validation that maximized AUC. This maximization of the AUC ensures that our operationalized versions of the manual scoring systems provide an upper bound on the performance the scoring system is capable of.

Table 6 and Fig. 9 show the performance for each operationalized scoring system on the screening task.

The manual version of some of the scoring systems we operationalized have previously been evaluated on the task of screening for general dementia. Results reported for Shulman ranged from 0.66 (Storey et al. 2002) to 0.79 (Storey et al. 2001), while our operationalization of Shulman yielded 0.67 on memory impairment disorders and 0.71 on vascular cognitive disorders. Results reported for Mendez ranged from 0.70 (Storey et al. 2002) to 0.78 (Storey et al. 2001), while our operationalization of Mendez gave us 0.72 on memory impairment disorders and 0.70 on vascular cognitive disorders. Manos achieved 0.67 (Lourenço et al. 2008), while our operationalization gave us 0.73 on memory impairment disorders and 0.69 on vascular cognitive disorders. Thus, while there is a range of accuracies reported for these algorithms due in part to their being evaluated on different datasets and for different groupings of conditions (general dementia vs. memory impairment disorders/vascular cognitive disorders), our operationalized scoring systems achieve similar accuracies, providing a check on our operationalization process.

We then used a variety of machine learning methods on the op-clinician features and the simplest features. The lower part of Table 6 shows AUCs for the best machine learning algorithm on these two feature sets, followed by the AUCs of the best machine learning

Table 5 Operationalization of Rouleau scoring system

Maximum: 10 points
1. Integrity of the clockface (maximum: 2 points)
2: eccentricity $\leq \epsilon_1$ AND clockface closed percentage $\geq \epsilon_2$
1: eccentricity $> \epsilon_1$ OR clockface closed percentage $< \epsilon_2$
No clockface strokes OR normed residual $> \epsilon_3$
2. Presence and sequencing of the numbers (maximum: 4 points)
4: If all digits present AND correct angular sequence AND DigitsAngleError $\leq \epsilon_4$
3: If all digits present AND correct angular sequence AND $\epsilon_4 \leq$ DigitsAngleError $\leq \epsilon_5$
2: (At least one digit missing OR at least one digit repeated OR digits greater than 12 present)
AND DigitNeighborsTest $\leq \epsilon_6$)
OR numbers counterclockwise
OR All digits present AND (at least one digit outside the clock OR DigitNeighborsTest $\geq \epsilon_6$)
1: At least one digit missing OR at least one digit repeated OR digits greater than 12 present)
AND DigitNeighborsTest $\geq \epsilon_6$)
No digits
3. Presence and placement of the hands (maximum: 4 points)
4: Exactly two hands AND both HandAngleError $\leq \epsilon_7$ AND HandRatio $\leq \epsilon_8$
3: Exactly two hands AND (at least one hand has $\epsilon_7 <$ HandAngleError $\leq \epsilon_9$ OR HandRatio $> \epsilon_8$)
2: Exactly two hands AND at least one hand has HandAngleError $> \epsilon_9$
2: OR Minute hand pointing closer to "10" than "2" and within 30° of digit "10"
1: One hand or more than two hands present
0: No hands present

algorithm on all features (reproduced from Sect. 4 for comparison). We can see that all three machine learning models are much more accurate than the operationalized scoring systems, even when using identical features (the op-clinician features), or ones that are even easier to measure (the simplest features).

Table 7 and Fig. 10 show corresponding accuracy results for the operationalized scoring systems on the diagnosis task. Again, the machine learning classifiers created from all three feature sets are much more accurate than the operationalized scoring systems, which scored mostly in the low 60s. We were unable to find any published accuracies for these existing scoring systems on a comparable diagnosis task. Given these higher accuracies from the machine learning models, the dCDT could be considered not only as a general screening tool, but could also potentially guide diagnosis.

6 Interpretable linear models

We have found that state-of-the-art machine learning methods on simplest features, clinician features, and the set of all features outperform existing scoring criteria. But the existing

Table 6 Operationalized scoring system AUCs for screening test, together with AUCs of the best machine learning model on the op-clinician features, simplest features, and the set of all features

Algorithm	Memory impairment disorders versus HC	Vascular cognitive disorders versus HC	PD versus HC	All three versus HC
Manos	0.73 (0.08)	0.69 (0.13)	0.70 (0.11)	0.70 (0.07)
Royall	0.73 (0.14)	0.67 (0.13)	0.73 (0.09)	0.70 (0.06)
Shulman	0.67 (0.05)	0.71 (0.07)	0.66 (0.07)	0.67 (0.05)
Libon	0.67 (0.09)	0.72 (0.09)	0.68 (0.10)	0.68 (0.12)
Rouleau	0.61 (0.16)	0.68 (0.15)	0.59 (0.13)	0.61 (0.08)
Mendez	0.72 (0.11)	0.70 (0.12)	0.69 (0.07)	0.69 (0.06)
MiniCog	0.57 (0.08)	0.55 (0.13)	0.54 (0.15)	0.58 (0.12)
Best ML with op-clinician features	0.83 (0.09)	0.83 (0.11)	0.86 (0.08)	0.82 (0.10)
Best ML with simplest features	0.83 (0.06)	0.82 (0.07)	0.83 (0.08)	0.83 (0.07)
Best ML with all features	0.93 (0.09)	0.88 (0.11)	0.91 (0.11)	0.91 (0.09)

F-scores are in "Appendix 2", Table 20

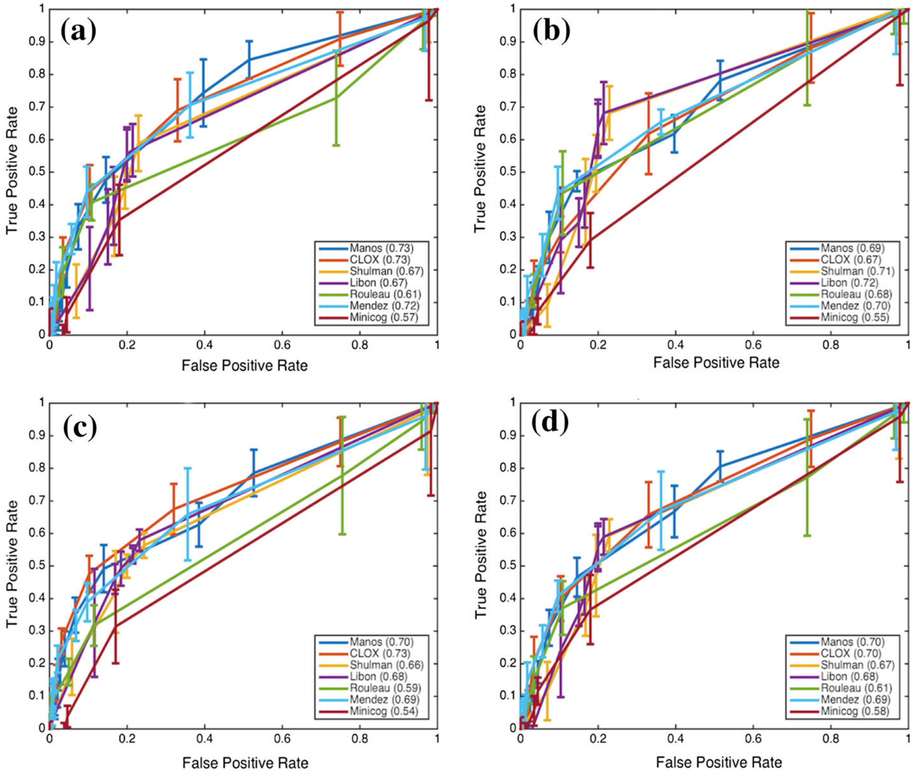

Fig. 9 ROC curves for the experiments in Table 6. **a** Memory impairment disorders versus HC, **b** vascular cognitive disorders versus HC, **c** PD versus HC, **d** all three versus HC

Table 7 Operationalized scoring system AUCs for diagnosis task, together with AUCs of the best machine learning model on the op-clinician features, simplest features, and the set of all features

Algorithm	Memory impairment disorders versus all others	Vascular cognitive disorders versus all others	PD versus all others	All three versus all others
Manos	0.69 (0.07)	0.63 (0.08)	0.62 (0.07)	0.64 (0.06)
Royall	0.68 (0.08)	0.62 (0.07)	0.65 (0.07)	0.63 (0.09)
Shulman	0.62 (0.07)	0.65 (0.05)	0.59 (0.06)	0.63 (0.04)
Libon	0.60 (0.08)	0.65 (0.12)	0.60 (0.14)	0.64 (0.05)
Rouleau	0.59 (0.13)	0.64 (0.09)	0.53 (0.09)	0.60 (0.06)
Mendez	0.68 (0.06)	0.65 (0.05)	0.61 (0.07)	0.61 (0.07)
MiniCog	0.55 (0.07)	0.56 (0.07)	0.53 (0.05)	0.54 (0.07)
Best ML with op-clinician features	0.73 (0.06)	0.71 (0.08)	0.71 (0.05)	0.70 (0.06)
Best ML with simplest features	0.72 (0.05)	0.73 (0.07)	0.74 (0.08)	0.72 (0.05)
Best ML with all features	0.83 (0.06)	0.79 (0.05)	0.82 (0.05)	0.82 (0.05)

F-scores are in “Appendix 2”, Table 21

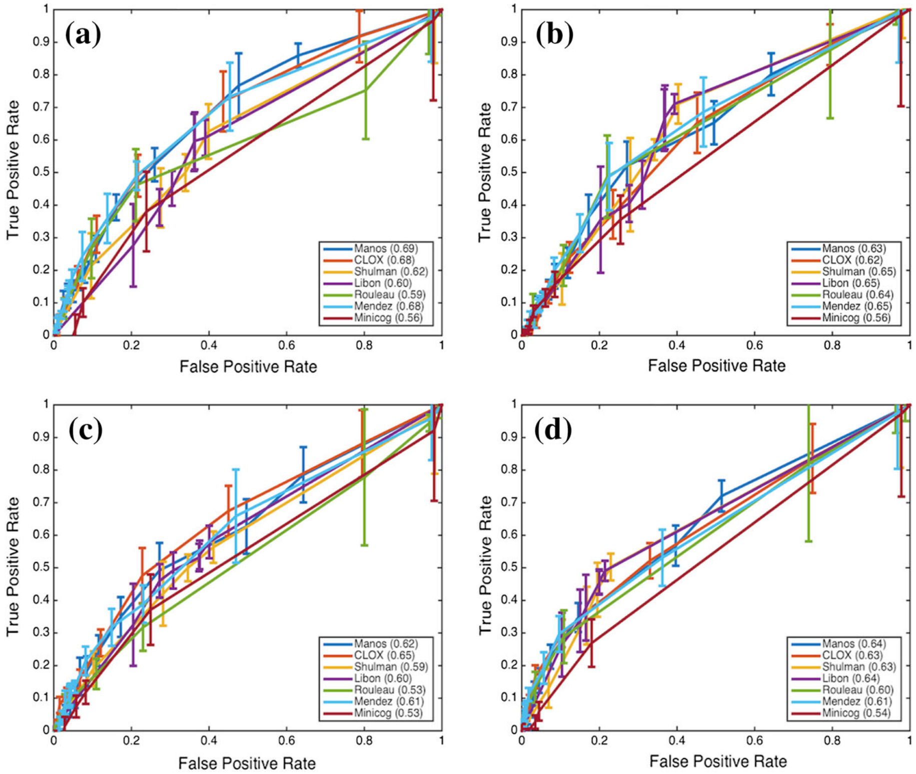

Fig. 10 ROC curves for the experiments in Table 7. **a** Memory impairment disorders versus all others, **b** vascular cognitive disorders versus all others, **c** PD versus all others, **d** all three versus all others

scoring systems remain more interpretable. Interpretability is crucial if domain experts are to accept and use the model. We turn next to finding models that are more transparent and hence more likely to be accepted in practice, yet still outperform existing models.

The interpretability of a model is domain specific. To ensure that we produced models that can be used and accepted in a clinical context, we obtained guidelines from clinicians. This led us to focus on three components: ease of feature measurements and their reliability, model computational complexity, and model understandability.

1. Ease of feature measurements and reliability: Some features can be measured quickly by eye (e.g. is there a minute hand present) while others would require a digital pen (time to draw the hand). In addition, some have a greater inter-clinician variance in measurements. This led us to focus on features that we believed would have the lowest variance; as noted we call these the "simplest features." Models produced using these features could easily be used even without a digital pen or other digitizing mechanism.
2. Computational complexity: the models should be relatively easy to compute, requiring a number of simple operations similar to the existing manual scoring systems. The existing scoring systems discussed above have on average 8–15 rules, with each rule containing on average one or two features. We thus focus on models that use fewer than 20 features, and have a simple form, which in our case means either addition or subtraction of feature scores (i.e., a linear model), or an ordered sequence of if-then statements (a rule list or decision list). Clinicians should be able to evaluate these types of models rapidly.
3. Understandability: the rationale for a decision made by the model should be easily understandable, so that the user can understand why the prediction was made and can easily explain it. Thus if several features are roughly equally useful in the model, the most understandable one should be used. As one example of what we mean by "understandable," note that our feature set includes 3 measures of test taking time: the total time to draw the command clock, the total time to draw the copy clock, and the aggregate of the two, the total time to draw both. If using total time to draw both clocks produces the most accurate model, but almost all of the predictive power comes from only one of the components, say the total time to draw the command clock, it would be reasonable to trade some small amount of accuracy in order to use the simpler feature, the command clock drawing time. The form of the model is also important for understandability, leading us to focus on linear models and rule lists.

Our goal in the remainder of this paper is to build classifiers that are at least as interpretable as existing scoring systems (according to the criteria mentioned above), but that are more accurate. While our focus will be on using the simplest features, we will also create interpretable models using op-clinician features and the MRMR subset of all features. These latter two might not be as practical to use manually, and may not be as interpretable, but exploring them allows us to test the predictive power of these more complex features. In addition, if these models achieve high accuracy, they could also be used for automatic scoring while providing interpretability for each prediction.

We begin by using a recently developed framework, Supersparse Linear Interpretable Models (SLIM) (Ustun and Rudin 2015; Ustun et al. 2013), designed to create sparse linear models that have integer coefficients and constraints on the range of coefficients. To improve model understandability, we added feature preferences, where certain features would be preferred over others if performance is similar.

Given a dataset of N examples $D_N = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$, where observation $\boldsymbol{x}_i \in \mathrm{R}^J$ and label $y_i \in \{-1, 1\}$, and an extra element with value 1 is included within each $\mathbf{x}_i$ vector to act as the intercept term, we want to build models of the form $\hat{y} = \text{sign}(\boldsymbol{\lambda}^T \boldsymbol{x})$, where $\boldsymbol{\lambda} \subseteq \mathbb{Z}^{J+1}$ is a vector of integer coefficients. The framework determines the coefficients of the models by solving an optimization problem of the form:

$$\begin{aligned} \min_{\boldsymbol{\lambda}} \quad & \text{Loss}(\boldsymbol{\lambda}; D_N) + \cdot \Phi(\boldsymbol{\lambda}) \\ \text{s.t.} \quad & \boldsymbol{\lambda} \in \mathcal{L}. \end{aligned}$$

The Loss function $\text{Loss}(\boldsymbol{\lambda}; D_N)$ penalizes misclassifications. The interpretability penalty function $\Phi(\boldsymbol{\lambda})\colon \mathbb{R}^{J+1} \to \mathbb{R}$ allows for a tradeoff between desired interpretability qualities and accuracy. The framework also allows interpretability constraints by limiting $\boldsymbol{\lambda}$ to a user-defined set $\mathcal{L}$, to restrict coefficients to a particular set of values, in our case, integers.

The interpretability penalty function $\Phi(\boldsymbol{\lambda})$ allows us to prioritize certain features, helping to ensure that the most understandable features appear in the model. In particular, we defined an understandability penalty u_j for each feature j by organizing our features into trees such that the children of each feature are those it depends on. For instance "total time to draw both clocks" has as children "total time to draw command clock" and "total time to draw copy clock." The height of a given node is the number of nodes traversed from the top of the tree to the given node. We define

$$u_j = \text{height}(j) \quad \forall j$$

which produces a bias toward simpler features, i.e., those lower in the tree. To regulate both the model complexity and the model understandability, we define our interpretability penalty function $\Phi(\boldsymbol{\lambda})$ as

$$\begin{aligned} \Phi(\boldsymbol{\lambda}) &= \text{sparsity penalty} + \text{understandability penalty} \\ &= C_0 \sum_{j=1}^{J} \mathbf{1}[\boldsymbol{\lambda}_j \neq 0] + C_1 \sum_{j=1}^{J} u_j \cdot \mathbf{1}[\boldsymbol{\lambda}_j \neq 0]. \end{aligned} \tag{1}$$

The first term simply computes the ℓ_0 semi-norm of $\Phi(\boldsymbol{\lambda})$, which is the count of the number of nonzero features. This term encourages the model to use fewer features. The second term allows the optimization to potentially sacrifice some training accuracy to favor using features lower in the tree, which we believe will be more understandable. The constants C_0 and C_1 trade off between the two terms, so that if one cares more about sparsity, then C_0 would be set larger, and if one cares more about understandability, then C_1 would be set larger. The values of C_0 and C_1 can be set using nested cross-validation if desired.

The loss function measures a balance between accuracy on the positive examples and accuracy on the negative examples. It is:

$$\text{Loss}(\boldsymbol{\lambda}; D_N) = C_+ \frac{1}{N} \sum_{i:y_i=1} \psi_i + C_- \frac{1}{N} \sum_{i:y_i=-1} \psi_i,$$

where ψ_i is 1 if an incorrect prediction is made. The user-defined values of C_+ and C_- determine the relative costs of false negatives and false positives.

We include a margin γ, and we say an incorrect prediction is made if the value of $y_i \boldsymbol{\lambda}^T \boldsymbol{x_i}$ is below γ.

The SLIM optimization thus becomes:

$$
\begin{aligned}
\min_{\boldsymbol{\lambda},\psi,\boldsymbol{\Phi},\alpha,\beta} \; & \frac{C_+}{N}\sum_{i:y_i=1}\psi_i + \frac{C_-}{N}\sum_{i:y_i=-1}\psi_i + \sum_{j=1}^{J}\Phi_j \\
\text{s.t.} \quad & M_i\psi_i \ge \gamma - \sum_{j=0}^{J} y_i\lambda_j x_{i,j} && i=1,\ldots,N && \textit{0–1 loss} && \text{(2a)}\\
& \Phi_j = C_0\alpha_j + C_1 u_j\alpha_j + \epsilon\beta_j && j=1,\ldots,J && \textit{int. penalty} && \text{(2b)}\\
& -\Lambda_j\alpha_j \le \lambda_j \le \Lambda_j\alpha_j && j=1,\ldots,J && \ell_0\textit{-norm} && \text{(2c)}\\
& -\beta_j \le \lambda_j \le \beta_j && j=1,\ldots,J && \ell_1\textit{-norm} && \text{(2d)}\\
& \lambda_j \in \mathcal{L}_j && j=0,\ldots,J && \textit{coefficient set} &&\\
& \psi_i \in \{0,1\} && i=1,\ldots,N && \textit{loss variables} &&\\
& \Phi_j \in \mathbb{R}_+ && j=1,\ldots,J && \textit{penalty variables} &&\\
& \alpha_j \in \{0,1\} && j=1,\ldots,J && \ell_0\textit{variables} &&\\
& \beta_j \in \mathbb{R}_+ && j=1,\ldots,J && \ell_1\textit{variables} &&
\end{aligned}
$$

The first constraints (2a) force ψ_i to be 1 if the prediction is incorrect, meaning below the margin value γ. The value of M_i is a sufficiently large upper bound on values of $\gamma - \sum_{j=0}^{J} y_i\lambda_j x_{i,j}$; it can be computed easily given the largest value of each $x_{i,j}$ and the largest allowable value of λ_j (denoted Λ_j): we set M_i such that $M_i > \gamma + \sum_j \Lambda_j \max_i(x_{i,j})$. The constraints (2b) define the two penalty terms in Φ and also include a very small coefficient on β_j, where β_j will be set to the absolute value of coefficient λ_j. The β_j term is not a regularization term, its only effect is to ensure that the coefficients are coprime, meaning that they are not divisible by a common integer. This helps with interpretability but does not affect training accuracy. For instance, consider coefficients [2,2,4] and [1,1,2], which are indistinguishable to all other terms in the model. The $\epsilon\beta_j$ terms will force the optimization to choose [1,1,2]. The next constraints (2c) define each α_j as being 1 when λ_j is non-zero (0 otherwise). Λ_j is the largest allowable value of λ_j. The constraints (2d) on β_j, along with the fact that β_j is being minimized in the objective, define it as being the absolute value of λ_j. The meaning of the other constraints was explained above.

The optimization problem was programmed in Matlab and solved using the CPLEX 12.6 API. We ran our optimization problem on the set of simplest features and the clinician features, with a hard upper bound of 10 features, to keep them interpretable, and on the MRMR subset of all features with an upper bound of 20 features. We used stratified cross-validation to divide the data into 5 folds to obtain training and testing sets. We further cross-validated each training set into 5 folds to optimize the parameters (C_+, C_-, C_0, C_1) using grid search over a set of ranges. Tables 8 and 9 present the AUCs for screening and diagnosis, respectively. For screening, all the SLIM models outperformed the operationalized scoring systems, the best of which performed in the 0.70–0.73 range (Table 8). For diagnosis, only the SLIM models with the MRMR subset of all features significantly outperforms the operationalized scoring systems, while the others perform similarly, the best of the operationalized systems performed in the 0.64–0.69 range (Table 9).

Table 10 shows a SLIM model containing only 9 binary features, yet it achieves an AUC score of 0.78. Pushed by the understandability penalty, the model uses mostly simple features composed of a single property, except for the first line which consists of an aggregate of multiple simpler features, chosen by the optimization despite its complexity because of its

Table 8 AUC results for Supersparse Linear Integer Models on the screening task

Features versus HC	Memory impairment disorders versus HC	Vascular cognitive disorders versus HC	PD versus HC	All three versus HC
SLIM with simplest features	0.78 (0.08)	0.75 (0.05)	0.78 (0.07)	0.74 (0.05)
SLIM with op-clinician features	0.75 (0.10)	0.74 (0.07)	0.73 (0.11)	0.74 (0.06)
SLIM with MRMR subset	0.83 (0.09)	0.81 (0.13)	0.81 (0.10)	0.83 (0.09)
Best operationalized scoring system	0.73 (0.08)	0.72 (0.09)	0.73 (0.09)	0.70 (0.06)
Best ML with all features	0.93 (0.09)	0.88 (0.11)	0.91 (0.11)	0.91 (0.09)
Best ML with op-clinician features	0.83 (0.09)	0.83 (0.11)	0.86 (0.08)	0.82 (0.10)
Best ML with simplest features	0.83 (0.06)	0.82 (0.07)	0.83 (0.08)	0.83 (0.07)

F-scores are in “Appendix 2”, Table 22

Table 9 AUC results for Supersparse Linear Integer Models on the diagnosis task

Features	Memory impairment disorders versus all others	Vascular cognitive disorders versus all others	PD versus all others	All three versus all others
SLIM with simplest features	0.68 (0.12)	0.66 (0.10)	0.66 (0.07)	0.69 (0.05)
SLIM with op-clinician features	0.67 (0.09)	0.66 (0.07)	0.66 (0.10)	0.70 (0.04)
SLIM with MRMR subset	0.75 (0.04)	0.72 (0.06)	0.77 (0.06)	0.76 (0.08)
Best operationalized scoring system	0.69 (0.07)	0.65 (0.05)	0.65 (0.07)	0.64 (0.05)
Best ML with all features	0.83 (0.06)	0.79 (0.05)	0.82 (0.05)	0.82 (0.05)
Best ML with op-clinician features	0.73 (0.06)	0.71 (0.08)	0.71 (0.05)	0.70 (0.06)
Best ML with simplest features	0.72 (0.05)	0.73 (0.07)	0.74 (0.08)	0.72 (0.05)

F-scores are in “Appendix 2”, Table 23

Table 10 Supersparse Linear Integer Model for screening of memory impairment disorders

PREDICT MEMORY IMPAIRMENT DISORDER IF SCORE < 10	
Command clock	
1. All digits are present, not repeated, and in the correct angular order	+5
2. Hour hand is present	+5
3. All of the non-anchor digits are in the correct eighth	+1
4. Crossed-out digits present	−3
5. Two hands not present	−1
6. More than 60 s to draw	−1
7. Minute hand points to digit 10	−6
Copy clock	
8. All of the non-anchor digits are in the correct eighth	+4
9. Numbers are repeated	−3

high screening power. This model contains only elements from the simplest feature set, which means they do not have the problems present in many existing scoring systems; in particular, the features used in the model are not as subjective, producing a scoring system likely to be more reliable.

7 Rules and rule lists

We mined association rules from our data and used these rules to build interpretable rule lists. The rules allow us to gain insights about how different cognitive impairments influence behavior on the test. By constraining the width and length of our decision lists to levels similar to existing scoring systems, and by using simple features, we created rule lists that we believe can be easily interpreted by clinicians. Unlike the linear models above, rules and rule lists also allow us to use non-linear relationships in the data.

7.1 Mining association rules

The first step was to discretize all of our features into equal-frequency bins, using 2 and 5 bins per feature. We then mined globally for all IF-THEN rules in the data that obeyed certain conditions on the quality of the rule. In particular, we wanted the rules with both sufficiently high support (i.e the number of subjects that obeyed the IF condition) and high confidence (i.e. the empirical probability of the THEN condition to be true, given that the IF condition is true). We used FPGrowth (Borgelt 2005) to extract decision rules from our data that predict each of our conditions (memory impairment disorders, vascular cognitive disorders, PD). We set a minimum support threshold of 40 tests, and required confidence to be greater than chance, where chance is simply the proportion of total patients who had the condition. Figure 11 shows the distribution of confidence and support for rules for each condition in the screening task.

These graphs show us that some of these rules can be very accurate. For memory impairment disorders for example, we have a rule that, for our data, can be applied to 15 % of the tests and can accurately predict memory impairment disorders over 80 % of the time (circled

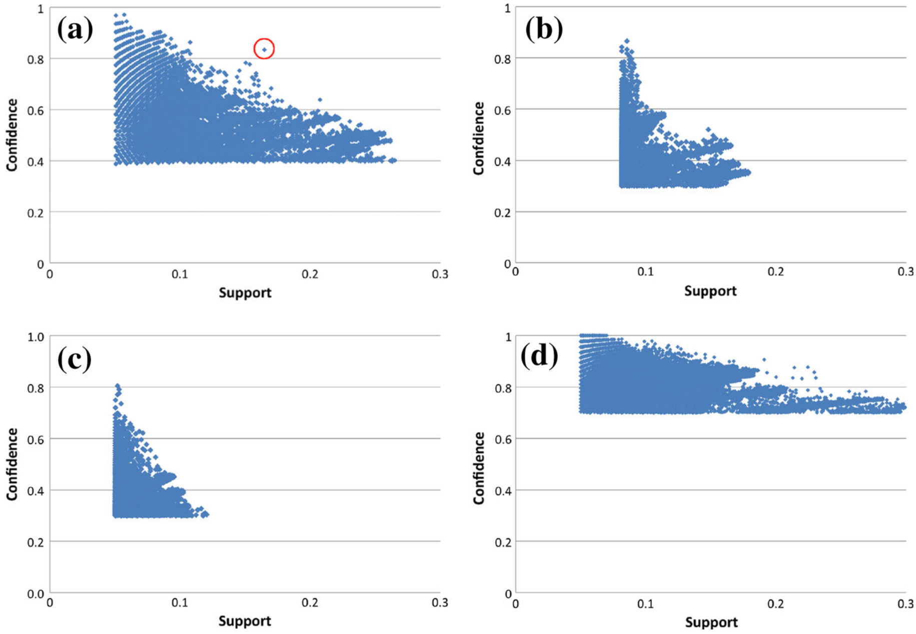

Fig. 11 Scatter plot of confidence versus support for rules for each condition versus HC. Each *dot* on the plot represents an IF-THEN rule, where the condition is the THEN part of the rule. The right angle at the bottom left of each of these clusters shows the minimum confidence and support cutoffs used when mining the rules. **a** Memory impairment disorders versus HC, **b** vascular cognitive disorders versus HC, **c** PD versus HC, **d** all three versus HC

in Fig. 11a). This rule is: Predict memory impairment if the pre-first-hand latency on the copy clock is greater than 2.3 s, and at least one hand is missing on the command clock. This rule is consistent with what is known about memory impairment disorders.

7.2 Interesting patterns

Some of the association rules confirm existing knowledge about correlations between pen-based features and clinical groups (e.g., the example just above). Others appear to be novel, possibly providing insight into correlations not reported previously. Tables 11, 12, and 13 present a set of rules that focus on the screening task for memory impairment disorders, vascular cognitive disorders, and PD.

7.2.1 Memory impairment disorders

The first rule in Table 11 shows that, when compared to HC subjects, the memory impairment group subjects tend to spend a greater percentage of the test-taking time thinking (i.e., with pen off the paper) and a smaller percentage of their test-taking time inking (with pen on the paper). This is consistent with what is known about Alzheimer's disease and amnestic MCI.

The second rule indicates that memory impairment group subjects make a longer than normal pause between the first stroke of the hands and the last stroke that was drawn before the hands on the copy clock. This may result from decision-making difficulty, or from trouble recalling the instructions given (e.g., what time to set the clock to). Combining this second

Table 11 Screening conditions implying memory impairment disorders

Rule	Support	Confidence
1. Percentage thinking time is high, >65 % (alternative phrasing: percentage inking time is low, <35 %)	0.2	0.57
2. Pre-first-hand latency on copy clock is high, >2.3 s	0.2	0.64
3. Pre-first-hand latency on copy clock is high, >2.3 s, AND at least one hand missing on the command clock	0.14	0.84
4. There is at least one digit missing on command clock and none missing on copy clock	0.04	0.78
5. The minute hand is pointing more than 15° away from digit 2 on command clock but points within 15° degrees on copy clock	0.06	0.75

Table 12 Screening conditions implying vascular cognitive disorders

	Rule	Support	Confidence
1	In the command clock the minute hand points within 15° of digit 10	0.04	0.79
2	In the command clock one or more digits have fewer than 2 other digits within ±45°	0.19	0.52
3	For both clocks average time to draw digits is high, >2.5 s	0.2	0.52

Table 13 Screening conditions implying PD

	Rule	Support	Confidence
1	Average inking time over both clocks is high, >17 s	0.2	0.38
2	Average angle gap over both clock faces is high, >57°	0.2	0.43
3	The average pen speed is low for both clocks	0.19	0.41
4	Average digit width over both clocks is low, <3 mm	0.2	0.33
5	Average digit height over both clocks is low, <5 mm	0.2	0.34
6	Average number of strokes per clock is high, >27	0.16	0.34
7	Average number of noise strokes per clock is high, >1.5	0.2	0.38
8	Average number of noise strokes smaller than 0.3 mm per clock is high, >0.5	0.2	0.49

rule with the condition that at least one hand is missing in the command clock gives the third rule, which has a very high confidence.

Memory impairment patients tend to display signs of improvement from the command clock to the copy clock. Consistent with this, the fourth rule finds in the data that there is a significant chance someone belongs in the memory impairment group if they have one or more digits missing on their command clock but none missing on their copy clock. Similarly, the fifth rule tells us that this group is very likely the correct choice if the minute hand is not aimed accurately in the command clock but is aimed accurately in the copy clock.

7.2.2 Vascular cognitive disorders

The patterns that distinguish the vascular-related cognitive disorders subjects from our HC subjects are similar to those of the memory impairment group. These subjects also tend to spend more time thinking, less time inking, and show signs of improvements between the two clocks.

We highlight a few additional rules in Table 12. The first rule shows a particularly interesting phenomenon: some patients draw the minute hand pointing towards the 10 digit instead of towards the 2 digit, presumably driven by the words "ten" and "eleven" (as in the instructions to set the time to "ten past eleven"). Almost 80 % of people who do this fall in our vascular cognitive disorders group, making it a very accurate rule for screening. The second rule measures the angular distribution of the digits around the clock, and if one or more digits have fewer than 2 other digits within ±45°, there is a high chance the subject belongs in our vascular cognitive disorders group. These subjects also tend to spend a long time drawing digits, as shown in the third rule.

7.2.3 Parkinson's disease

The patterns for the Parkinson's group are very different. As expected, given the motor slowing and increased incidence of tremor characteristic of this disorder, instead of having low inking time like the memory group and the cognitive disorders group, subjects in the PD group tend to have high inking time over both clocks, likely due to motor impairment, as shown in the first rule of Table 13. The second rule shows that they tend to leave a larger angular gap in their clock faces, possibly a consequence of their difficulty in starting, stopping, and persisting in motions, which might contribute to premature stopping, producing the gaps. They also tend to display signs of bradykensia, drawing slower than HC patients, a common symptom of Parkinson's, as shown in the third rule. The fourth and fifth rule show that the digits tend to be both shorter and narrower than those of HC subjects, suggestive of micrographia, also common among Parkinson's patients. Both their command and copy clocks also tend to have more total strokes (rule 6), and they also have a larger number of noise strokes (rule 7), particularly small strokes (rule 8), possibly due to tremors, or a pull to stimulus (i.e. the subject is resting the pen on a target of attention in the clock).

While all the rules described above provide interesting insights when considered individually, we also want to combine them to produce a classifier in the form of a rule list, yielding a classifier with a high degree of accuracy that remains interpretable. We turn next to this.

7.3 Rule lists

To construct scoring systems for the CDT that are both accurate and interpretable using the rules mined above, we chose a recently developed machine learning algorithm called *Bayesian Rule Lists* (BRL) (Letham et al. 2015). Its intent is to create classifiers that have better accuracy and interpretability than traditional machine learning models like CART, and thus more likely to be used by clinicians. BRL derives from the data an ordered list of IF-THEN rules. Table 16 shows an example of a BRL list for screening of memory impairment disorders.

There are two main steps to the BRL algorithm:

– Find all of the feature combinations that occur sufficiently often (e.g., copy clock is missing numbers AND there is a missing hour hand on the command clock).

- Choose and order the feature combinations to form the left hand sides of rules for the rule list. This is done using a Bayesian modeling approach. BRL has two user-defined parameters that enter into its Bayesian prior over rule lists, allowing the user to specify the desired number of rules in the rule list, λ, and the desired number of conditions within each rule, η.

BRL's Bayesian modeling approach creates a posterior distribution of rule lists. The Bayesian prior encourages it to favor lists with approximately λ rules and η conditions per rule, as specified by the user.

We ran BRL on our three sets of features: simplest features, op-clinician features, and the MRMR subset of all features. AUCs for screening are shown in Table 14, and range from 0.79 to 0.85. These are significantly more accurate than the operationalized scoring systems (the best of which performed in the 0.70–0.73 range, Table 6). These scores are approximately at the level of the best ML algorithms with simplest features and with op-clinician features. Diagnosis AUCs, shown in Table 15, display a range from 0.69 to 0.74, again better than the operationalized scoring systems (the best of which performed in the 0.64–0.69 range, Table 7), and similar to the best ML algorithms with simplest features and with op-clinician features.

To keep the rule lists interpretable, we restricted the width of the lists to at most 2. We varied the list length to see how it influences accuracy. Figure 12 shows testing AUC versus list length for the simplest features. For the screening task, Fig. 12 indicates that lists between 4 and 7 rules lead to test AUCs similar to the best test AUC's, while for the testing task, 5–8 rules leads to diagnoses similar to the highest AUC's. These models are generally both more concise and more accurate than the existing scoring algorithms we discussed earlier.

Table 16 presents a rule list obtained for the screening of memory impairment disorders. This rule list was derived using the simplest features to allow the resulting rule list to be used with the pen-and-paper test, and to allow clinicians to measure these features quickly and reliably by eye. Containing only 5 rules, each of similar complexity to a line from the existing scoring systems, it is shorter than most of the existing scoring systems, yet it achieves an AUC of 0.82, higher than the upper bound of 0.73 on the best existing scoring system that we examined.

Another new machine learning method, similar to Bayesian Rule Lists, is called Falling Rule lists (Wang and Rudin 2015). A falling rule list is a rule list where the right hand side probabilities decrease along the list. This means that the most at-risk patients should fall into one of the first few conditions. This can be useful for decision-making, as clinicians need only check the first few conditions on the list to determine whether a patient is at high risk. We ran FRL on the set of simplest features, and obtained the model in Table 17 for the screening of all three conditions versus healthy; it contains only five rules yet has an AUC of 0.75.

8 Discussion of challenges and conclusion

Traditional scoring systems created by clinicians are typically based on obvious features and thus have a transparency and face validity that is readily understood by the user population. A potential lack of transparency in machine learning-derived classifiers could be a barrier to clinical use.

Our goal was to have the best of both worlds: create an automated system that took advantage of new sensor technology (the digital pen), state-of-the-art machine learning methods,

Table 14 AUC results for BRL on screening task

Features	Memory impairment disorders versus HC	Vascular cognitive disorders versus HC	PD versus HC	All three versus HC
BRL with simplest features	0.82 (0.06)	0.79 (0.08)	0.81 (0.05)	0.82 (0.06)
BRL with op-clinician features	0.82 (0.07)	0.78 (0.07)	0.83 (0.09)	0.78 (0.10)
BRL with MRMR subset	0.83 (0.10)	0.82 (0.07)	0.79 (0.09)	0.85 (0.09)
Best operationalized scoring system	0.73 (0.08)	0.72 (0.09)	0.73 (0.09)	0.70 (0.06)
Best ML with all features	0.93 (0.09)	0.88 (0.11)	0.91 (0.11)	0.91 (0.09)
Best ML with op-clinician features	0.83 (0.09)	0.83 (0.11)	0.86 (0.08)	0.82 (0.10)
Best ML with simplest features	0.83 (0.06)	0.82 (0.07)	0.83 (0.08)	0.83 (0.07)

F-scores are in “Appendix 2”, Table 24

Table 15 AUC results for BRL on diagnosis task

Features	Memory impairment disorders versus all others	Vascular cognitive disorders versus all others	PD versus all others	All three versus all others
BRL with simplest features	0.72 (0.08)	0.71 (0.05)	0.70 (0.08)	0.69 (0.06)
BRL with op-clinician features	0.70 (0.11)	0.72 (0.08)	0.69 (0.07)	0.72 (0.11)
BRL with MRMR subset	0.73 (0.08)	0.70 (0.05)	0.73 (0.06)	0.74 (0.08)
Best operationalized scoring system	0.69 (0.07)	0.65 (0.05)	0.65 (0.07)	0.64 (0.05)
Best ML with all features	0.83 (0.06)	0.79 (0.05)	0.82 (0.05)	0.82 (0.05)
Best ML with op-clinician features	0.73 (0.06)	0.71 (0.08)	0.71 (0.05)	0.70 (0.06)
Best ML with simplest features	0.72 (0.05)	0.73 (0.07)	0.74 (0.08)	0.72 (0.05)

F-scores are in “Appendix 2”, Table 25

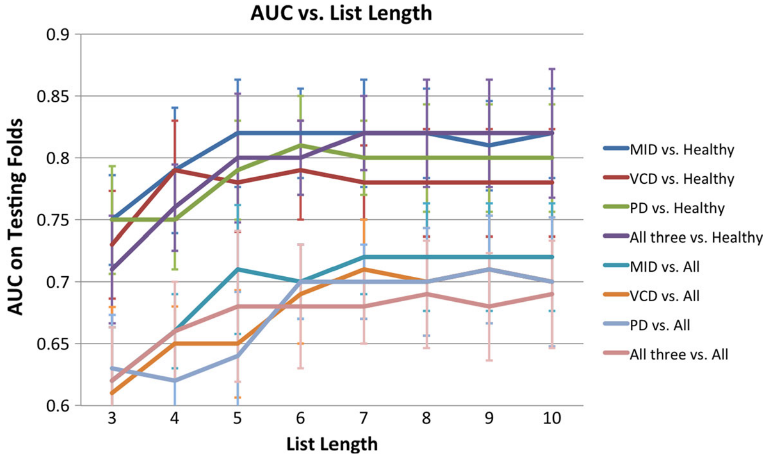

Fig. 12 Plot of AUC on testing folds versus list length for simplest features, for both screening and diagnosis

Table 16 BRL for screening of memory impairment disorders

IF the command clock minute hand points within 15° of digit 10	**THEN** 94 % (88–100 %)
ELSE IF the command clock minute hand is present and drawn outwards from the center **AND** all of the non-anchor digits in the command clock are in the correct eighth	**THEN** 16 % (12–20 %)
ELSE IF all hands are present with arrowheads pointing outwards **AND** more than 5 of the non-anchor digits in the copy clock are in the correct eighth	**THEN** 24 % (17–32 %)
ELSE IF the command clock took more than 40 s to draw	**THEN** 92 % (84–98 %)
ELSE IF the copy clock took less than 20 s to draw	**THEN** 12 % (0–21 %)
ELSE	33 % (12–45 %)

Percentages are the probability of memory impairment disorders, with the 95 % confidence interval in parentheses

and large amounts of patient data, but that ensured the same interpretability qualities as the existing scoring systems. There are several important challenges we faced when trying to create our assessment models, in addition to the usual challenges of applying machine learning in practice for knowledge discovery applications.

The first challenge is *interpretability*. A major theme of this work is how to walk the line between interpretability and accuracy. We started with traditional (black box) machine learning methods to establish the highest accuracy baselines, then went to the other end of the spectrum by mining association rules, which provided accuracy baselines for the most interpretable methods. We then aimed to find the right balance of interpretability and accuracy using new machine learning techniques designed for this particular tradeoff. The models we learned have major advantages in accuracy over the traditional scoring systems for the clock

Table 17 FRL for screening of memory impairment disorders, vascular cognitive disorders, and Parkinson's disease

IF not all non-anchor digits in the command clock are in the correct eighth **AND** there is at least one hand missing	**THEN** 88 %
ELSE IF not all non-anchor digits in the copy clock are in the correct eighth **AND** the copy clock took more than 30 s to draw	**THEN** 68 %
ELSE IF there is at least one noise stroke in the copy clock **AND** there are fewer than 4 definite hooklets in the copy clock	**THEN** 65 %
ELSE IF there are more than two digits an the incorrect quadrant in the copy clock **AND** the copy clock took more than 20 s to draw	**THEN** 34 %
ELSE	15 %

Percentages are the probability of any of the three disorders

drawing test, and even some advantages in interpretability because the traditional pen-and-paper scoring systems require subjective judgment and are not consistent across clinicians. Interpretability is notoriously difficult to quantify for a particular domain, but in this case we were able to use new machine learning techniques to create models that mimic the form of model that clinicians currently use. These techniques allowed us to optimize directly for interpretability as we chose to define it. The resulting models are potentially directly usable by clinicians. Our results indicate that some of our models are more robust, just as interpretable, more accurate than some widely used scoring systems, and require less computation on the part of the clinicians to compute the result, even without the benefit of the detailed data from the digital pen.

Another challenge we faced is how to create a reasonable assessment of the quality of our predictions, which required us to encode subjective human judgments in a way that captured the intent of those judgments. This led to our strategy of creating an optimized version of each of the existing scoring systems (the operationalized scoring systems). We were then able to show that even fully optimized versions of widely used scoring methods were not as accurate as a machine learning methods trained on data—even when that machine learning method was trained on the same features used in the existing scoring systems. This shows the power of combining machine learning with clinical knowledge.

This project brings together many important pieces: a new sensor (the digital pen), new techniques for handwritten stroke classification, techniques for optimizing calculations made using human judgment, new machine learning techniques for interpretability, and data created from many subjects' clock drawings and their subsequent clinical classifications. While our classifiers now need to be tested in actual clinical use, the results presented here suggest the potential of this work to make significant improvements in both screening and diagnosis.

Acknowledgments This work has been supported in part by: the Robert E. Wise Research and Education Institution, Defense Advanced Research Projects Agency Contract D13AP00008, National Science Foundation Award IIS-1404494, National Institute of Neurological Disorders and Stroke Grants R01-NS17950, K23-NS60660, and R01-NS082386, National Heart, Lung, and Blood Institute Contract N01-HC25195, National

Institutes on Aging Grants R01 AG0333040, AG16492, and AG08122, National Institute on Mental Health Grant RO1-MH073989, IH/NCATS Clinical and Translational Science Award to the University of Florida UL1TR000064, and the University of Florida Movement Disorders and Neurorestoration. We thank the Anoto Corporation, which from the outset of this project has generously provided technical advice and guidance.

Appendix 1: Machine learning implementation details

CART The R library “rpart” with default parameters.

C4.5 The R library “RWeka”” with default settings.

SVM: SVMlight Joachims (1998) with a radial basis function kernel. We selected the slack parameter C_{SVM} and the kernel parameter γ using a grid search over the ranges $C_{SVM} \in \{2^{-4}, 2^{-2}, \ldots, 2^{14}\}$ and $\gamma \in \{2^{-6}, 2^{-1}, \ldots, 2^{10}\}$

Random Forests The MATLAB class “TreeBagger” with parameter “NVarToSample” set to the square root of the total number of variables and the variable “NTrees” for the number of trees was set to 1000.

Regularized Logistic Regression The LIBLINEAR Fan et al. (2008) implementation of logistic regression with l_1 regularization. We selected the regularization parameter C_{LR} from $\{2^{-8}, 2^{-6}, \ldots, 2^{8}\}$ as that with the best 5-fold cross-validation performance.

Boosted Decision Trees The MATLAB class “fitensemble” with 500 trees and parameter “LearnRate” set to 0.05.

Appendix 2: F-scores of experiments

Table 18 Machine learning algorithms F-scores for screening test

Algorithm	Memory impairment disorders versus HC	Vascular cognitive disorders versus HC	PD versus HC	All three versus HC
C4.5	0.67 (0.07)	0.57 (0.08)	0.58 (0.05)	0.77 (0.08)
CART	0.72 (0.09)	0.63 (0.06)	0.58 (0.09)	0.76 (0.05)
SVM Gaussian	0.79 (0.04)	0.69 (0.04)	0.65 (0.06)	0.86 (0.08)
Random forest	0.77 (0.03)	0.74 (0.05)	0.74 (0.06)	0.83 (0.05)
Boosted decision trees	0.82 (0.04)	0.73 (0.03)	0.67 (0.04)	0.85 (0.05)
Regularized logistic regression	0.77 (0.02)	0.69 (0.04)	0.74 (0.03)	0.83 (0.05)

Table 19 Machine learning algorithms F-scores for diagnosis test

Algorithm	Memory impairment disorders versus all others	Vascular cognitive disorders versus all others	PD versus all others	All three versus all others
C4.5	0.38 (0.05)	0.34 (0.04)	0.35 (0.07)	0.41 (0.05)
CART	0.41 (0.05)	0.37 (0.04)	0.34 (0.04)	0.40 (0.07)
SVM Gaussian	0.51 (0.04)	0.59 (0.03)	0.54 (0.04)	0.48 (0.03)
Random forest	0.53 (0.02)	0.48 (0.05)	0.51 (0.04)	0.47 (0.02)
Boosted decision Trees	0.49 (0.03)	0.49 (0.06)	0.5 (0.04)	0.57 (0.04)
Regularized logistic regression	0.51 (0.04)	0.52 (0.03)	0.53 (0.02)	0.52 (0.04)

Table 20 Operationalized scoring system F-scores for screening test

Algorithm	Memory impairment disorders versus HC	Vascular cognitive disorders versus HC	PD versus HC	All three versus HC
Manos	0.59 (0.04)	0.48 (0.07)	0.51 (0.05)	0.71 (0.03)
Royall	0.59 (0.04)	0.45 (0.03)	0.53 (0.05)	0.70 (0.06)
Shulman	0.58 (0.03)	0.52 (0.08)	0.48 (0.06)	0.69 (0.05)
Libon	0.57 (0.06)	0.53 (0.03)	0.50 (0.05)	0.68 (0.04)
Rouleau	0.50 (0.04)	0.48 (0.03)	0.38 (0.03)	0.67 (0.05)
Mendez	0.58 (0.02)	0.50 (0.07)	0.47 (0.08)	0.69 (0.04)
MiniCog	0.50 (0.05)	0.37 (0.03)	0.38 (0.03)	0.69 (0.05)

Table 21 Operationalized scoring system F-scores for diagnosis test

Algorithm	Memory impairment disorders versus all others	Vascular cognitive disorders versus all others	PD versus all others	All three versus all others
Manos	0.27 (0.03)	0.17 (0.02)	0.17 (0.02)	0.39 (0.04)
Royall	0.28 (0.05)	0.14 (0.05)	0.18 (0.07)	0.37 (0.08)
Shulman	0.23 (0.06)	0.17 (0.07)	0.14 (0.03)	0.42 (0.06)
Libon	0.24 (0.10)	0.17 (0.08)	0.16 (0.04)	0.43 (0.07)
Rouleau	0.26 (0.04)	0.19 (0.02)	0.13 (0.04)	0.35 (0.03)
Mendez	0.28 (0.09)	0.18 (0.07)	0.17 (0.11)	0.37 (0.05)
MiniCog	0.21 (0.03)	0.13 (0.02)	0.14 (0.04)	0.34 (0.06)

Table 22 F-scores results for Supersparse Linear Integer Models on screening test

Features	Memory impairment disorders versus HC	Vascular cognitive disorders versus HC	PD versus HC	All three versus HC
SLIM with simplest features	0.66 (0.04)	0.63 (0.04)	0.61 (0.03)	0.72 (0.05)
SLIM with op-clinician features	0.64 (0.02)	0.58 (0.06)	0.57 (0.04)	0.73 (0.07)
SLIM with MRMR subset	0.72 (0.05)	0.65 (0.06)	0.63 (0.02)	0. 78 (0.02)

Table 23 F-scores results for Supersparse Linear Integer Models on diagnosis test

Features	Memory impairment disorders versus all others	Vascular cognitive disorders versus all others	PD versus all others	All three versus all others
SLIM with simplest features	0.34 (0.04)	0.33 (0.02)	0.32 (0.07)	0.42 (0.05)
SLIM with op-clinician features	0.32 (0.05)	0.34 (0.05)	0.32 (0.04)	0.43 (0.05)
SLIM with MRMR subset	0.46 (0.04)	0.47 (0.06)	0.49 (0.03)	0.47 (0.03)

Table 24 F-scores results for BRL on screening test

Features	Memory impairment disorders versus HC	Vascular cognitive disorders versus HC	PD versus HC	All three versus HC
BRL with simplest features	0.74 (0.03)	0.64 (0.02)	0.60 (0.05)	0.79 (0.06)
BRL with op-clinician features	0.72 (0.04)	0.64 (0.02)	0.63 (0.03)	0.76 (0.02)
BRL with MRMR subset	0.74 (0.03)	0.66 (0.01)	0.61 (0.04)	0.81 (0.03)

Table 25 F-scores results for BRL on diagnosis test

Features	Memory impairment disorders versus all others	Vascular cognitive disorders versus all others	PD versus all others	All three versus all others
BRL with simplest features	0.40 (0.05)	0.39 (0.02)	0.36 (0.07)	0.45 (0.03)
BRL with op-clinician features	0.37 (0.3)	0.42 (0.04)	0.36 (0.04)	0.46 (0.03)
BRL with MRMR subset	0.43 (0.03)	0.41 (0.05)	0.41 (0.06)	0.49 (0.05)

Appendix 3: All operationalized scoring systems

Additional features

We define two additional non-obvious features that appear within the operationalized scoring systems, in Table 26. The following subsections each provide an existing scoring system and our operationalization of it.

Manos

Table 27 provides the original Manos scoring system, and Table 28 shows our operationalization.

Royall

Table 29 provides the original Royall scoring system, and Table 30 shows our operationalization.

Shulman

Table 31 provides the original Schulman scoring system, and Table 32 shows our operationalization.

Table 26 Additional non-obvious operationalized clinician features

Variable	Description
ClockfaceGap	The distance between the start and end of the clock face
DigitClockfaceDistanceVariance	The variance in the distance of digits from the clockface
HandIntersectCenterDistance	The distance between where the hands intersect (or would intersect) and the center of the best fit ellipse, normalized for the size of the ellipse

Table 27 Original Manos scoring system (Manos and Wu 1994)

Maximum: 10 points

1. Digit placement errors (maximum: 8 points)

The clock is divided into eighths, beginning with a line through “12” and

the center of the circle

(if “12” is missing the position is assumed to be counterclockwise from the “1”

at a distance equal to that between the “1” and “2”)

For each eighths, add one point if the expected anchor digit is missing

2. Presence and placement of the hands (maximum: 2 points)

One point each is given for an obvious short hand pointing at the “11”

and an obvious long hand pointing to the “2”

The difference in the length of the hands must be obvious at a glance

Table 28 Operationalization of Manos scoring system

```
Maximum: 10 points
1. Digit placement errors (maximum: 8 points)
  Get angle of digit 12
    If “12” present, go to step 2.
    Else if “12” not present but “1” and “2” present, get angle of “1” and “2”, compute difference in angle,
    and add difference to angle of “1” to get approximate angle of “12”.
    Else if “12” not present but “10” and “11”” present, get angle of “10” and “11”, compute difference
    in angle, and subtract difference to angle of “11” to get approximate angle of “12”.
    Else, bring up error.
$\forall$ step $\in [-15, -14, \ldots, 0, \ldots, 14, 15]$
    Break up clock into eighths using angle of “12” + step
    and adding multiples of 45° to obtain eighths
    For each eighth, add one point if the expected anchor digit is missing
Pick the minimum score over all step values.
2. Presence and placement of the hands (maximum: 2 points)
  If exactly two hands are present AND handRatio $\leq \epsilon_1$
    If minute hand has handAngleError $\leq \epsilon_2$, add 1
    If hour hand has handAngleError $\leq \epsilon_2$, add 1
```

Table 29 Original Royall scoring system (Royall et al. 1998)

Maximum: 15 points; one point for each line satisfied

1. Does figure resemble a clock?
2. Circular face present?
3. Dimensions >1 inch?
4. All numbers inside the perimeter?
5. “12”, “6”, “3” and “9” placed first?
6. Spacing intact? (symmetry on either side of “12” and “6” o’clock)
7. No sectoring or tic marks?
8. Only Arab numerals?
9. Only numbers 1–12 among the numerals present?
10. Sequence 1–12 intact? (no omissions or intrusions)
11. Only two hands present? (ignore sectoring/tic marks)
12. All hands represented as arrows?
13. Hour hand between 1 and 2 o’clock?
14. Minute hand longer than hour hand?
15. None of the following

    (1) hand pointing to 10 o’clock

    (2) “11:10” present?

    (3) intrusions from “hand” or “face” present?

    (4) any letters, words or pictures?

    (5) any intrusion from circle below?

Table 30 Operationalization of Royall scoring system

Maximum: 15 points; one point for each line satisfied

1. Clockface closed percentage $\geq \epsilon_1$ AND at least 4 digits present AND at least 1 hand present

2. Clockface present

3. Major axis of fitted ellipse to clockface greater than 1 inch

4. All numbers inside the clockface

5. "12", "6", "3", "9" all anchor digits

6. DigitsAngleError $\leq \epsilon_2$

7. No spokes or tick marks present

8. Always 1 (we do not have any clocks with other numerals in our dataset so assume it is very rare)

9. No digit greater than 12 present

10. All numbers present in correct order by angle, no repetitions,
    no numbers greater than 12, no text, crossed-out digits allowed

11. Two hands present, no repetitions of hands but allow crossed-out hands

12. Arrows present on both hands, direction must be correct

13. Angle of hour hand between angle of "11" and angle of "12"

If either digits or hand missing, 0

14. HandRatio $\leq \epsilon_3$

15. None of the following

(1) Minute hand pointing closer to "10" than "2" and within 30° of digit "10"

(2) Any text present

(3) Always false. Very hard to measure, and no example in dataset so assume it is very rare

(4) Any text present

(5) Always false

Table 31 Original Shulman scoring system (Shulman et al. 1993)

Maximum: 6 points

1. Perfect

2. Minor visuospatial errors

Examples

(a) Mildly impaired spacing of times

(b) Draws times outside circle

(c) Turns page while writing numbers so that some numbers appear upside down

(d) Draws in lines (spokes) to orient spacing

3. Inaccurate representation of “10 after 11” when visuospatial organization is perfect or shows only minor deviations

Examples

(a) Minute hand points to “10”

(b) Writes “10 after 11”

(c) Unable to make any denotation of time

4. Moderate visuospatial disorganization of times such that accurate denotation of “10 after 11” is impossible

Example

(a) Moderately poor spacing

(b) Omits numbers

(c) Perseveration: repeats circle or continues on past 12–13, 14, 15 etc.

(d) Right-left reversal: numbers drawn counterclockwise

(e) Dysgraphia: unable to write numbers accurately

5. Severe level of disorganization as described in 4

6. No reasonable representation of a clock

Exclude severe depression or other psychotic states

Example

(a) No attempt at all

(b) No semblance of a clock at all

(c) Writes a word or name

Table 32 Operationalization of Shulman scoring system

Maximum: 6 points

1. eccentricity $\leq \epsilon_1$ AND clockface closed percentage $\geq \epsilon_2$ AND

all digits present AND no digits repeated AND correct angular sequence AND DigitsAngleError $\leq \epsilon_3$ AND

exactly two hands AND both have HandAngleError $\leq \epsilon_4$

2. Minor visuospatial errors

(a) $\epsilon_3 <$ DigitsAngleError $\leq \epsilon_5$

(b) At least one digit outside the circle

(c) No way to measure automatically given our data, and very rare according to doctors

(d) At least one spoke present

3. Inaccurate representation of "10 after 11" when visuospatial organization is perfect or shows only minor deviations

(a) Minute hand pointing closer to "10" than "2" and within 30° of digit "10"

(b) Any text present

(c) both hands have HandAngleError $> \epsilon_4$

4. Moderate visuospatial disorganization of times such that accurate denotation of "10 after 11" is impossible

(a) DigitNeighborsTest $\geq \epsilon_6$

(b) At least one digit missing

(c) More than one clockface OR at least one digit repeated OR digits greater than 12 present

(d) Numbers drawn counterclockwise

(e) At least one digit missing

5. Severe level of disorganization as described in 4

Severely poor spacing: DigitNeighborsTest $\geq \epsilon_6$

6. No reasonable representation of a clock

Clockface closed percentage $< \epsilon_2$ OR fewer than four digits present OR no hands present

Libon

Table 33 provides the original Libon scoring system, and Table 34 shows our operationalization.

Table 33 Original Libon scoring system (Libon et al. 1993)

Maximum: 10 points

Scores 10–6: Circle and Hands are basically intact, some impairment in hand placement

10: Hands, numbers and circle are totally intact

9: Slight error(s) in hand number placement; hands of equal length; any self-correction

8: More noticeable errors in hand/number placement; hand length correct but shifted to one side or top/bottom

7: Significant errors in hand placement; hand placement intact with some numbers deleted; minor perseveration in number placement

6: Inappropriate use of clock hands i.e., digital display; circling numbers to indicate hand placement; connecting the numbers 10 and 11 or 11 and 2

Scores 5–1: Circle, numbers and/or hand placement are grossly impaired

5: Crowding numbers to one side; numbers reversed; significant perseveration of numbers within circle boundary

4: Loss of clock face integrity, numbers outside circle boundary, further distortion of number placement

3: Numbers and clock face no longer connected

2: Vague representations of a clock; clock face absent but numbers present

1: Either no attempt or response is made; scattered bits or fragments are produced

Table 34 Operationalization of Libon scoring system

Maximum: 10 points

Scores 10–6: Circle and Hands are basically intact, some impairment in hand placement

10: Both hands have HandAngleError $\leq \epsilon_1$

9: At least one hand has $\epsilon_1 <$ HandAngleError

8: Both hands hand have $\epsilon_1 <$ HandAngleError

7: At least one hand not in correct quadrant

6: Ignore: Hard to measure automatically, and very rare in our data

Scores 5–1: Circle, numbers and/or hand placement are grossly impaired

5: DigitNeighborsTest $\geq \epsilon_2$

4: Any number missing or any number placed outside the clockface

3: DigitNeighborsTest $\geq \epsilon_3$

2: Clockface closed percentage $< \epsilon_4$ OR fewer than four digits present OR less than one hand present

1: Clockface closed percentage $< \epsilon_4$ AND fewer than four digits present AND less than one hand present

Mendez

Table 35 provides the original Mendez scoring system, and Table 36 shows our operationalization.

Table 35 Original Mendez scoring system (Mendez et al. 1992)

Maximum: 20 points; one point for each line satisfied

1. There is an attempt to indicate a time in any way
2. All marks or items can be classified as either part of a closure figure, a hand, or a symbol for clock numbers
3. There is a totally closed figure without gaps (closure figure)

Score only if symbols for clock numbers are present

4. A 2 is present and is pointed out in some way for the time
5. Most symbols are distributed as a circle without major gaps
6. Three or more clock quadrants have one or more appropriate numbers
   (12–3, 3–6, 6–9, 9–12 per respective clockwise quadrant)
7. Most symbols are ordered in a clockwise or rightward direction
8. All symbols are totally within a closure figure
9. An 11 is present and is pointed out in some way for the time
10. All numbers 1–12 are indicated
11. There are not repeated or duplicated number symbols
12. There are no substitutions for Arabic or Roman numerals
13. The numbers do not go beyond the number 12
14. All symbols lie about equally adjacent to a closure figure edge
15. Seven or more of the same symbol type are ordered sequentially.

Score only if one or more hands are present

16. All hands radiate from the direction of a closure figure center
17. One hand is visibly longer than another hand
18. There are exactly two distinct and separable hands
19. All hands are totally within a closure figure
20. There is an attempt to indicate a time with one or more hands.

Table 36 Operationalization of Mendez scoring system

Maximum: 20 points; one point for each line satisfied

1. At least one hand present
2. No noise, no ticks, no spokes, no text
3. ClockfaceGap $\leq \epsilon_1$

Score only if symbols for clock numbers are present

4. "2" is present, minute hand has handAngleError $\leq \epsilon_2$
5. DigitsAngleError $< \epsilon_3$
6. Break clock into quadrants, and at least three correct digits within each quadrant
7. More than half of digits present are in clockwise direction
8. No digit or hands present outside clockface
9. "11" is present, hour hand has handAngleError $\leq \epsilon_2$
10. All digits present
11. No repeated digits (cross-outs allowed)
12. Always 1 (hard to measure and very rare in our data)
13. No digits greater than 12 present
14. DigitClockfaceDistanceVariance $< \epsilon_4$
15. At least 7 digits are in correct order by angle

Score only if one or more hands are present

16. HandIntersectCenterDistance $\leq \epsilon_5$
17. HandRatio $\leq \epsilon_6$
18. Only two hands present (cross-outs allowed)
19. Ignore since never happens in our data
20. At least one hand present

MiniCog

Table 37 provides the original MiniCog scoring system, and Table 38 shows our operationalization.

Table 37 Original MiniCog scoring system (Borson et al. 2000)

Maximum: 1 point	
If all numbers approximately in the correct position AND there are two hands pointing properly	+1

Table 38 Operationalization of MiniCog scoring system

Maximum: 1 point	
If DigitsAngleError $< \epsilon_1$ AND both hands have HandAngleError $< \epsilon_2$	+1

Appendix 4: Grid search for operationalization parameters

See Table 39.

Table 39 Parameter search values for operationalizations

Variable	Threshold values
Eccentricity of fitted ellipse	{0.5, 0.53, . . . , 0.77, 0.80}
ClockfaceClosedPercentage	{50, 55, . . . , 95, 100}
ClockfaceGap	{0, 0.3, . . . , 2.7, 3.0}
DigitsAngleError	{0, 10, . . . , 90, 100}
DigitNeighborsTest	{0, 1, 2, 3, 4}
DigitClockfaceDistanceVariance	{0.1, 0.2, . . . , 0.9, 1.0}
HandAngleError	{0, 2, . . . , 38, 40}
HandRatio	{0.7, 0.75, . . . , 1.0, 1.05}

References

ABS Consulting. (2002). *Marine safety: Tools for risk-based decision making*. New York: Rowman & Littlefield.

Albert, M. S., DeKosky, S. T., Dickson, D., Dubois, B., Feldman, H. H., Fox, N. C., et al. (2011). 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. *Alzheimers Dement*, *7*(3), 270–279.

Andrade, J. T. (2009). *Handbook of violence risk assessment and treatment: New approaches for mental health professionals*. Berlin: Springer.

Battistin, L., & Cagnin, A. (2010). Vascular cognitive disorder. A biological and clinical overview. *Neurochemical Research*, *35*(12), 1933–1938.

Borgelt, C. (2005). An implementation of the FP-growth algorithm. In *Proceedings of the 1st international workshop on open source data mining: Frequent pattern mining implementations, OSDM '05* (pp. 1–5).

Borson, S., Scanlan, J., Brush, M., Vitaliano, P., & Dokmak, A. (2000). The mini-cog: a cognitive 'vital signs' measure for dementia screening in multi-lingual elderly. *International Journal of Geriatric Psychiatry*, *15*(11), 1021–1027.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Belmont: Wadsworth.

Cohen, J., Penney, D. L., Davis, R., Libon, D. J., Swenson, R. A., Ajilore, O., et al. (2014). Digital clock drawing: Differentiating 'thinking' versus 'doing' in younger and older adults with depression. *Journal of the International Neuropsychological Society*, *20*(09), 920–928.

Davis, R., & Penney, D.L. (2014). *Method and apparatus for measuring representational motions in a medical context*. US Patent 8,740,819.

Davis, R., Penney, D., Pittman, D., Libon, D., Swenson, R., & Kaplan, E. (2011). *The Digital Clock Drawing Test (dCDT)—I: Development of a new computerized quantitative system*. Presented at the 39th annual meeting of the International Neuropsychological Society, Boston, MA.

Davis, R., Libon, D. J., Au, R., Pitman, D., & Penney, D. L. (2014). THink: Inferring cognitive status from subtle behaviors. In *Twenty-sixth IAAI conference* (pp. 2898–2905).

Fan, R. E., Chang, K. W., Hsieh, C. J., Wang, X. R., & Lin, C. J. (2008). LIBLINEAR: A library for large linear classification. *The Journal of Machine Learning Research*, *9*, 1871–1874.

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

Freitas, A. A., Wieser, D. C., & Apweiler, R. (2010). On the importance of comprehensible classification models for protein function prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, *7*(1), 172–182.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189–1232.

Grande, L., Rudolph, J., Davis, R., Penney, D., Price, C., & Swenson, R. (2013). Clock drawing: Standing the test of time. In Ashendorf Le Swenson (Ed.), *The Boston process approach to neuropsychological assessment*. Oxford: Oxford University Press.

Haury, A. C., Gestraud, P., & Vert, J. P. (2011). The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. *PloS ONE*, *6*(12), e28,210.

Hauser, J. R., Toubia, O., Evgeniou, T., Befurt, R., & Dzyabura, D. (2010). Disjunctions of conjunctions, cognitive simplicity, and consideration sets. *Journal of Marketing Research*, *47*(3), 485–496.

Joachims, T. (1998). Making large-scale SVM learning practical. LS8-report 24, Universität Dortmund, LS VIII-report.

Kim, H., Cho, Y. S., & Do, E. Y. L. (2011a). Computational clock drawing analysis for cognitive impairment screening. In *Proceedings of the fifth international conference on tangible, embedded, and embodied interaction*. ACM (pp. 297–300).

Kim, H., Cho, Y. S., & Do, E. Y. L. (2011b). Using pen-based computing in technology for health. *Human–computer interaction. Users and applications* (pp. 192–201). Berlin: Springer.

Lamar, M., Grajewski, M., Penney, D., Davis, R., Libon, D., & Kumar, A. (2011). *The impact of vascular risk and depression on executive planning and production during graphomotor output across the lifespan*. Paper at 5th Congress of the International Society for Vascular, Cognitive and Behavioural Disorders, Lille, France.

Letham, B., Rudin, C., McCormick, T. H., & Madigan, D. (2015). Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model. *Annals of Applied Statistics* (accepted). http://imstat.org/aoas/next_issue.html.

Libon, D. J., Swenson, R. A., Barnoski, E. J., & Sands, L. P. (1993). Clock drawing as an assessment tool for dementia. *Archives of Clinical Neuropsychology*, *8*(5), 405–415.

Lourenço, R. A., Ribeiro-Filho, S. T., Moreira, Id F H, Paradela, E. M. P., & Miranda, A Sd. (2008). The Clock Drawing Test: Performance among elderly with low educational level. *Revista Brasileira de Psiquiatria*, *30*(4), 309–315.

Manos, P. J., & Wu, R. (1994). The ten point clock test: A quick screen and grading method for cognitive impairment in medical and surgical patients. *The International Journal of Psychiatry in Medicine*, *24*(3), 229–244.

Markatou, M., Tian, H., Biswas, S., & Hripcsak, G. (2005). Analysis of variance of cross-validation estimators of the generalization error. *The Journal of Machine Learning Research, 6,* 1127–1168.

Martens, D., Baesens, B., Van Gestel, T., & Vanthienen, J. (2007). Comprehensible credit scoring models using rule extraction from support vector machines. *European Journal of Operational Research*, *183*(3), 1466–1476.

Mendez, M. F., Ala, T., & Underwood, K. L. (1992). Development of scoring criteria for the clock drawing task in Alzheimer's disease. *Journal of the American Geriatrics Society*, *40*(11), 1095–1099.

Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., et al. (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., et al. (2013). The Framingham Heart Study clock drawing performance: Normative data from the offspring cohort. *Experimental Aging Research*, *39*(1), 80–108.

Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(8), 1226–1238.

Penney, D., Libon, D., Lamar, M., Price, C., Swenson, R., Eppig, J., et al. (2011a). *The Digital Clock Drawing Test (dCDT)—I: Information contained within the "noise", 5th Congress of the 1International Society for Vascular, Cognitive and Behavioural Disorders (VAS-COG)*, Lille, France.

Penney, D., Libon, D., Lamar, M., Price, C., Swenson, R., Scala, S., et al. (2011b). The Digital Clock Drawing Test (dCDT)—IV: Clock drawing time and hand placement latencies in mild cognitive impairment and dementia, abstract and poster. In *5th Congress of the International Society for Vascular, Cognitive and Behavioural Disorders*, Lille, France.

Penney, D., Lamar, M., Libon, D., Price, C., Swenson, R., Scala, S., et al. (2013). The Digital Clock Drawing Test (dCDT)—Hooklets: A novel graphomotor measure of executive function, abstract and poster. In

6th Congress of the International Society for Vascular, Cognitive and Behavioural Disorders, Montreal, Canada.

Petersen, R., Caracciolo, B., Brayne, C., Gauthier, S., Jelic, V., & Fratiglioni, L. (2014). Mild cognitive impairment: A concept in evolution. *Journal of Internal Medicine*, *275*(3), 214–228.

Plassman, B. L., Langa, K. M., Fisher, G. G., Heeringa, S. G., Weir, D. R., Ofstedal, M. B., et al. (2007). Prevalence of dementia in the United States: The aging, demographics, and memory study. *Neuroepidemiology*, *29*(1–2), 125–132.

Price, C. C., Cunningham, H., Coronado, N., Freedland, A., Cosentino, S., Penney, D. L., et al. (2011). Clock drawing in the Montreal Cognitive Assessment: Recommendations for dementia assessment. *Dementia and Geriatric Cognitive Disorders*, *31*(3), 179–187.

Prince, M., Guerchet, M., & Prina, M. (2013). *Policy brief for heads of government: The global impact of dementia 2013–2050*. London: Alzheimer Disease International.

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. Los Altos: Morgan Kaufmann.

Ridgeway, G. (2013). The pitfalls of prediction. *NIJ Journal, National Institute of Justice*, *271*, 34–40.

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

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

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

Steinhart, D. (2006). Juvenile detention risk assessment: A practice guide to juvenile detention reform. *Juvenile Detention Alternatives Initiative A project of the Annie E Casey Foundation*, *28*, 2011.

Storey, J. E., Rowland, J. T., Basic, D., & Conforti, D. A. (2001). A comparison of five clock scoring methods using ROC (receiver operating characteristic) curve analysis. *International Journal of Geriatric Psychiatry*, *16*(4), 394–399.

Storey, J. E., Rowland, J. T., Basic, D., & Conforti, D. A. (2002). Accuracy of the clock drawing test for detecting dementia in a multicultural sample of elderly Australian patients. *International Psychogeriatrics*, *14*(03), 259–271.

Strub, R. L., Black, F. W., & Strub, A. C. (1985). *The mental status examination in neurology*. Philadelphia: FA Davis.

Sun, H. (2006). An accurate and interpretable bayesian classification model for prediction of hERG liability. *ChemMedChem*, *1*(3), 315–322.

Sunderland, T., Hill, J. L., Mellow, A. M., Lawlor, B. A., Gundersheimer, J., Newhouse, P., et al. (1989). Clock drawing in Alzheimer's disease: A novel measure of dementia severity. *Journal of the American Geriatrics Society*, *37*(8), 725–729.

Tian, L., & Tibshirani, R. (2011). Adaptive index models for marker-based risk stratification. *Biostatistics*, *12*(1), 68–86.

Tuokko, H., Hadjistavropoulos, T., Rae, S., & O'Rourke, N. (2000). A comparison of alternative approaches to the scoring of clock drawing. *Archives of Clinical Neuropsychology*, *15*(2), 137–148.

Ustun, B., & Rudin, C. (2015). Supersparse linear integer models for optimized medical scoring systems. *Machine Learning*. doi:10.1007/s10994-015-5528-6.

Ustun, B., Tracà, S, & Rudin, C. (2013). Supersparse linear integer models for predictive scoring systems. In *Proceedings of AAAI late breaking track*

Van Belle, V. M., Van Calster, B., Timmerman, D., Bourne, T., Bottomley, C., Valentin, L., et al. (2012). A mathematical model for interpretable clinical decision support with applications in gynecology. *PloS ONE*, *7*(3), e34,312.

Verbeke, W., Martens, D., Mues, C., & Baesens, B. (2011). Building comprehensible customer churn prediction models with advanced rule induction techniques. *Expert Systems with Applications*, *38*(3), 2354–2364.

Wang, F., & Rudin, C. (2015). Falling rule lists. In *Proceedings of artificial intelligence and statistics (AISTATS)* (pp. 1013–1022).

Wang, T., Rudin, C., Doshi, F., Liu, Y., Klampfl, E., & MacNeille, P. (2015). *Bayesian Or's of And's for interpretable classification with application to context aware recommender systems* (submitted).