Journal of Alzheimer's Disease 70 (2019) 277–286
DOI 10.3233/JAD-190165
IOS Press

Episodic-Memory Performance in Machine Learning Modeling for Predicting Cognitive Health Status Classification

Michael F. Bergeron[a,*], Sara Landset[b], Franck Tarpin-Bernard[c], Curtis B. Ashford[d], Taghi M. Khoshgoftaar[b] and J. Wesson Ashford[e]
[a]*SIVOTEC Analytics, Boca Raton, FL, USA*
[b]*Department of Computer and Electrical Engineering and Computer Science, Florida Atlantic University, Boca Raton, FL, USA*
[c]*HAPPYneuron, S.A.S., Lyon, France*
[d]*MemTrax, LLC., Redwood City, CA, USA*
[e]*War-Related Illness and Injury Study Center, VA Palo Alto Health Care System and Department of Psychiatry & Behavioral Sciences, Stanford University School of Medicine, Palo Alto, CA, USA*

Handling Associate Editor: David Loewenstein

Accepted 1 May 2019

Abstract.
Background: Memory dysfunction is characteristic of aging and often attributed to Alzheimer's disease (AD). An easily administered tool for preliminary assessment of memory function and early AD detection would be integral in improving patient management.
Objective: Our primary aim was to utilize machine learning in determining initial viable models to serve as complementary instruments in demonstrating efficacy of the MemTrax online Continuous Recognition Tasks (M-CRT) test for episodic-memory screening and assessing cognitive impairment.
Methods: We used an existing dataset subset ($n = 18{,}395$) of demographic information, general health screening questions (addressing memory, sleep quality, medications, and medical conditions affecting thinking), and test results from a convenience sample of adults who took the M-CRT test. M-CRT performance and participant features were used as independent attributes: true positive/negative, percent responses/correct, response time, age, sex, and recent alcohol consumption. For predictive modeling, we used demographic information and test scores to predict binary classification of the health-related questions (yes/no) and general health status (healthy/unhealthy), based on the screening questions.
Results: ANOVA revealed significant differences among HealthQScore groups for response time true positive ($p = 0.000$) and true positive ($p = 0.020$), but none for true negative ($p = 0.0551$). Both %responses and %correct had significant differences ($p = 0.026$ and $p = 0.037$, respectively). Logistic regression was generally the top-performing learner with moderately robust prediction performance (AUC) for HealthQScore (0.648–0.680) and selected general health questions (0.713–0.769).
Conclusion: Our novel application of supervised machine learning and predictive modeling helps to demonstrate and validate cross-sectional utility of MemTrax in assessing early-stage cognitive impairment and general screening for AD.

Keywords: Aging, Alzheimer's disease, dementia, mass screening

*Correspondence to: Michael F. Bergeron, PhD, FACSM, SIVOTEC Analytics, Boca Raton Innovation Campus, 4800 T-Rex Avenue, Suite 315, Boca Raton, FL 33431, USA. E-mail: mbergeron@sivotecanalytics.com.

INTRODUCTION

Memory dysfunction is notably characteristic of aging and can often be attributed to Alzheimer's disease (AD) [1]. With its widespread prevalence and escalating incidence and public health burden [2], a simple tool that can be readily distributed and easily administered for valid preliminary assessment of memory function and early AD detection would be desirable and integral in improving patient management. Such advance insight could also be instrumental in potentially slowing the disease progression. Specifically, quick, clear, and valid insight into cognitive health status as an initial screen could measurably assist in diagnostic support and planning an individualized stratified care approach in medically managing those patients with early onset cognitive impairment. The computerized MemTrax tool (http://www.memtrax.com) was explicitly designed for such a purpose, and it is based on a simple and brief online and highly germane timed episodic memory challenge where the user responds to repeat images and not to any initial presentation [3, 4]. However, the clinical efficacy of this new approach in initial AD screening has not been sufficiently demonstrated or validated.

Traditional assessment of episodic memory using selected words recall or reproducing figures are characteristically imprecise, non-specific, and unreliable [5, 6]. And even more complex and contemporary computerized versions designed to address the multidimensional aspects of the memory process fail to measurably improve accuracy, reliability, or clinical interpretation across a highly variable spectrum of individual memory disorders and related subcomponents [7, 8]. These deficiencies in screening and detection remain barriers to suitably addressing the growing and widespread prevalence of AD and those affected [2].

There are numerous integrated and influencing factors to consider in interpreting the complex, highly variable, and evolving individual exhibiting characteristics of AD onset and progression. This presents a consequent well-recognized challenge to clinicians in validly assessing cognitive function and potential impairment, especially longitudinally. To better guide the practitioner in this difficult assessment and more optimally direct informed clinical management, advances in technology supported by artificial intelligence and machine learning could provide a distinct practical advantage. Notable examples featuring clinical utility of machine learning in brain health screening include Falcone et al. [9] who used Support Vector Machine (SVM) to detect concussion based on isolated vowel sounds extracted from speech recordings. Dabek and Caban [10] also utilized SVM in predictive modeling of military service members developing post-traumatic stress disorder after traumatic brain injury. And Climent et al. [11] conducted a cross-sectional study including an extensive array of clinically relevant variables and two screening tests while using decision tree machine learning modeling and complementary ensemble techniques to detect early mild cognitive impairment and associated risk factors in older adults. This new approach in utilizing machine learning to address the complexity of various human health challenges is only recent; but the demonstrated advantages in more aptly considering myriad interrelated factors that reflect the multiple domains of real-world systems biology are increasingly being realized. Accordingly, to thoroughly validate the practical clinical utility of MemTrax, individual test performance characteristics and a selected respective array of relevant influencing variables (e.g., age, medications, symptoms, etc.) must be considered and appropriately analyzed and modeled concomitantly in aggregate.

In this study, we explored an existing dataset consisting of demographic information, answers to general health screening questions (addressing memory, sleep quality, medications, and medical conditions affecting thinking), and test results from a convenience sample of adult individuals who took the MemTrax online Continuous Recognition Tasks (M-CRT) test for episodic-memory screening [3, 4]. We then performed predictive modeling on these data, using the demographic information and test scores to predict binary classification of the health-related questions (yes/no) and general health status. Thus, our primary aim was to utilize machine learning in determining initial viable models to serve as complementary instruments toward ultimately demonstrating the validated efficacy of MemTrax (via the M-CRT in this instance) as a clinical decision support screening tool for assessing cognitive impairment. Whereas the connection between responses to the general health-related questions and individual health status in the context of cognitive impairment was only speculative, we hypothesized that these self-reported indicators and the M-CRT online performance features would be confirmed as effective in our preliminary modeling to demonstrably support the low-cost and easily administered practical and relevant clinical efficacy of MemTrax.

MATERIALS AND METHODS

Data overview

The original dataset consisted of 30,435 instances of the M-CRT test conducted online between 9/22/11 and 8/23/2013 as part of the HAPPYneuron program (http://www.happy-neuron.com/) [3]. The study from which these data were provided for our present analysis was previously reviewed and approved by and administered in accord with the ethical standards of the Human Subject Protection Committee of Stanford University. The convenience sample was mix of people (adults) who were participating in this structured program to stimulate cognition. Whereas the sample was not truly representative of the general population, these individuals were generally healthy, though some may have had light cognitive or other impairments.

There were 25,146 total users who each took the test between 1 and 24 times. Each instance comprised 20 attributes including information from each user and respective test instance. The M-CRT online test included 50 images (25 unique and 25 repeats; 5 sets of 5 images of common scenes or objects) shown in a specific pseudo-random order. The participant would (per instructions) press the space bar of the computer to begin viewing the images series and again press the space bar as quickly as possible whenever a repeated picture appeared. Each image appeared for 3 s or until the space-bar was pressed, which prompted immediate (after 500 ms) presentation of the next picture. Response time was measured using the internal clock of the local computer and was recorded for every image, with a full 3 s recorded indicating no response. Response times less than 300 ms were interpreted as "no response". Additional details of the M-CRT administration and implementation, data reduction, and other data analyses are described elsewhere [3]. We focused our modeling on the four health-related screening questions and corresponding answers in the dataset. These questions were included in the M-CRT to establish, via self-reporting, whether each test respondent: 1) Has memory problems; 2) Has difficulty sleeping; 3) Is taking any medication; 4) Has any medical conditions that might affect his/her thinking.

Data cleaning

We first cleaned and examined the data for descriptive purposes and to determine the scope and incidence of information at hand. We followed a similar data cleaning process as described by Ashford et al. [3] to remove seemingly invalid M-CRT test results from the data prior to analysis. One criterion dictated eliminating M-CRT tests from users who provided invalid birth dates (indicating ages less than 21 years or over 99 years on the date of the test). Tests from users who did not provide their sex or who provided 5 or fewer total responses were also eliminated. This resulted in 18,477 tests from 18,395 users (based on unique user ID). With same-day and tests taken on subsequent days (after his/her first test) by the same user removed to eliminate bias from repeat instances and potential learning effects, we used only the 18,395 unique user tests for our analyses and health-related questions prediction modeling.

Data transformation

For our exploration, the data did not require an extensive amount of cleaning beyond the steps described above; but there were some additional items we addressed prior to beginning our analysis. Three attributes in the original dataset had responses in both English and French. Two of these attributes, occupation and employment status, were not used as part of our initial analysis, as they were not deemed relevant to our aims for this study; accordingly, these are not addressed/utilized here. For the third attribute regarding whether the user suffered from memory problems, the dataset was populated with values of "Yes," "No," "Oui," or "Non" (or left blank). Because this translation is unambiguous, we translated the French answers into English prior to completing our analysis.

The original data did not include the user's age; but we were able to derive ages by the user's birthday and date of the respective test, thus creating a numerical attribute representing the user's age on the date the M-CRT test was taken. For precision, age was represented in days rather than years in our analyses and models.

Based on the M-CRT test results, two derived features were created for each individual user's overall engagement: one for the percentage of total images shown to which the user registered an active response (keyboard spacebar click) and the other to indicate the percentage of the repeat and initial images (50 total) to which the user responded correctly. Percentage of total images prompting a response (%responses) was calculated using an established [3, 4] formula: true positive + (25 – true negative) with this total being divided by 50 representing the total images shown.

The percentage of correct responses (%correct) was calculated using the formula true positive+true negative divided by 50.

Finally, we created an additional new attribute called HealthQScore, so that we could quantify each user's collective answers to the four general health questions. Assigning each response to these questions a value of 0 or 1, all M-CRT test instances were given an aggregate HealthQScore between 0 and 4, based on the number of general health questions the user answered affirmatively. A HealthQScore was assigned only to test instances where the user provided answers to *all* four general health questions (and it was the user's first test, as repeat tests were already eliminated). Thus, we had a set of 4,645 M-CRT unique user tests from which to develop our general health status (HealthQScore) prediction models.

Experimental datasets

For these preliminary experiments, we created eight versions of the original data, using each of the individual general health questions, as well as various forms of the aggregate score, as the alternating dependent variable. Broadly, each derived dataset served one of two purposes: 1) Prediction of answers to individual general health questions or 2) Prediction of general health status based on HealthQScore. For each of the eight dataset versions, the following M-CRT performance and participant characteristic (demographic) features were used as independent attributes: true positive/negative, %responses/correct, response time true positive, age, sex, and whether the user had consumed alcohol in the preceding 24 h. For predictive modeling, we used the demographic information and test scores to predict binary classification of the health-related questions (yes/no) and general health status (healthy/unhealthy) for the test taker, based on the provided answers to the screening questions.

For each of the general health questions (memory problems, medications, difficulty sleeping, and medical conditions that affect thinking), two variations of each dataset were created, both using the respective general health question attribute as the class label. An instance was part of the positive class if the user answered "Yes" to the question, or it was part of the negative class if the user answered "No". In one variation of each of these datasets, the answers to the other three general health questions were used as independent features, while in the other variation they were not included. This distinction is denoted as 4Q and 1Q, respectively.

The four HealthQScore versions of the data differed based on how the data were split for binary classification. The binary classifications of the HealthQScore versions were based on our assumption that the more questions to which the user responded affirmatively, the more likely he or she is at risk for having a cognitive brain health deficit. Thus, we started our exploratory analysis by examining the most extreme cases only (scores of 0 versus 4 as the negative and positive classes, respectively) to see how well we could differentiate between the two groups. The challenge with this approach was that it only allowed for 1,004 instances to be used for analysis, which may not have been enough to build robust models on this dataset. For this reason, we also added combinations with aggregate scores of 1 and 3 into the negative and positive classes (0 or 1 versus 4; 0 or 1 versus 3 or 4; and 0 versus 3 or 4). All four (4) combinations of these groupings were tested.

A summary of all datasets and respective variations is presented in Table 1.

Descriptive statistics

Significant differences among HealthQScore groups for selected components of M-CRT

Table 1
Summary of datasets and variations (indicating respective # of participants) used for preliminary analysis

Dataset	Total Instances	Negative Class Size	Positive Class Size
MemoryProblems_1Q/4Q	17,042	6,822 (40.0%)	10,220 (60.0%)
Medications_1Q/4Q	4,999	2,854 (57.1%)	2,145 (42.9%)
DifficultySleeping_1Q/4Q	5,496	3,007 (54.7%)	2,489 (45.3%)
MedicalConditions_1Q/4Q	5,506	4,601 (83.6%)	905 (16.4%)
HealthQScore_0v4	1,004	679 (67.6%)	325 (32.4%)
HealthQScore_01v4	2,431	2,106 (86.6%)	325 (13.4%)
HealthQScore_01v34	3,203	2,106 (65.8%)	1,097 (34.3%)
HealthQScore_0v34	1,776	679 (38.2%)	1,097 (61.8%)

performance—i.e., True Positive, True Negative, %Responses, %Correct, and Response Time True Positive—were determined using Analysis of Variance (ANOVA). These same M-CRT performance metrics differentiated by answers to each of the general health questions were also compared using ANOVA.

Predictive modeling

For our preliminary analysis, we built 10 models for each of the 12 variations of our dataset to predict responses to the four general health questions and calculated index of general health status by binary classification. The 10 learners chosen for this analysis were 5-Nearest Neighbors (5NN), two versions of C4.5 decision tree (C4.5D and C4.5N), Logistic Regression (LR), Multilayer Perceptron (MLP), Naïve Bayes (NB), two versions of Random Forest (RF100 and RF500), Radial Basis Functional Network (RBF), and Support Vector Machine (SVM). Detailed descriptions to explain and contrast these algorithms have been described elsewhere [12]. These were chosen because they represent a variety of different types of learners and because we have had demonstrated success using these in previous experiments. Moreover, the parameter settings were chosen based on our previous research which showed them to be robust on a variety of different data [13]. Because this was a preliminary investigation and because our data were limited, further tuning of parameters was not employed as it would have increased the risk of overfitting our models and thus reduced the broader clinical utility beyond these specific data.

Each model was built using 10-fold cross validation, and model performance was measured using Area Under the ROC Curve (AUC). Our cross-validation process began with randomly dividing each of the 12 data sets into 10 equal segments, using nine of these respective segments to train the model and the remaining segment for testing. The number of instances in each segment varied by the size of the respective dataset as indicated in Table 1 (i.e., 1/10 of the total number of instances for each dataset) This procedure was repeated 10 times, using a different segment as the test set in each iteration. The results were then combined to calculate the final model's result/performance. For each learner/dataset combination, this entire process was repeated 10 times with the data being split differently each time. Repeating this procedure reduced bias, ensured replicability, and helped in determining the overall model performance. Differences between learner-specific model performance were examined using ANOVA and Tukey's Honest Significant Difference (HSD) test. In total, 12,000 models were built (12 datasets $\times$ 10 learners $\times$ 10 runs $\times$ 10 folds = 12,000 models).

RESULTS

Of our 18,395 test results, 17,405 included answers to at least one of the four general health questions (most users only answered one of these questions). The distribution of the number of answers for each question is shown in Table 2. Only 4,645 of the M-CRT participants included answers to all four general health questions.

Among the five available performance attributes to describe the M-CRT test results (true positive/negative, %responses/correct, and response time true positive), certain patterns emerged demonstrating an apparent link to a higher HealthQScore. Using a 95% confidence level, ANOVA revealed significant differences among HealthQScore groups for response time true positive ($p = 0.000$). There were also significant differences among HealthQScore groups for true positive ($p = 0.020$), but none for true negative ($p = 0.0551$). Both %responses and %correct also had significant differences ($p = 0.026$ and $p = 0.037$, respectively). Further examination showed that for both true positive and %responses, those with a HealthQScore of 0 performed significantly better than those with a 3 ($p = 0.0253$ and $p = 0.0166$, respectively), but all other HealthQScore groups (1, 2, and 4) overlapped with both. A similar pattern was demonstrated with %correct, as there were significant differences between participants with a HealthQScore of 1 and those with a 4 ($p = 0.0402$), but the other three groups (0, 2, and 3) overlapped with both. For the true positive response time variable, those respondents with a HealthQScore of 0 responded significantly faster than those with a 1 or 2 ($p = 0.000$), who in turn responded significantly faster than those with a HealthQScore of 3 or 4 ($p = 0.000$). Mean M-CRT test results for all five performance attributes

Table 2
Distribution of number of answers to each general health question

Question	# of answers	% of all instances
Memory Problems	17,042	92.6%
Medications	4,999	27.2%
Difficulty Sleeping	5,496	29.9%
Medical Conditions Affecting Thinking	5,506	29.9%

Table 3
Mean M-CRT test results separated by HealthQScore group

Score Group (# of instances)	True Positive*	True Negative	%Responses †	%Correct‡	Response Time True Positive§
0 (679)	21.1	22.8	46.6%	87.9%	901.4
1 (1427)	21.0	23.0	45.9%	87.9%	943.1
2 (1442)	20.9	22.8	46.2%	87.4%	960.0
3 (772)	20.6	23.0	45.2%	87.1%	1011.0
4 (325)	20.6	22.5	46.3%	86.2%	1023.2

*Significant differences among HealthQScore groups for True Positive (ANOVA; $p = 0.020$). †Significant differences among HealthQScore groups for %Responses (ANOVA; $p = 0.026$). ‡Significant differences among HealthQScore groups for %Correct (ANOVA; $p = 0.037$). §Significant differences among HealthQScore groups for Response Time True Positive (ANOVA; $p = 0.000$).

Table 4
Mean M-CRT test results differentiated by answers to general health questions

Question	Response	True Positive	True Negative	%Responses	%Correct	Response Time True Positive
Memory Problems	Yes	23.0	23.8	48.5%	93.6%	918.5
	No	23.4*	23.8	49.2%†	94.5%*	876.1‡
	n/a	23.3*	24.0	48.5%	94.5%*	888.1
Medications	Yes	20.8	22.8	46.1%	87.3%	998.7
	No	20.9	23.0	45.9%	87.7%	930.1
	n/a	24.0	24.2	49.8%	96.4%	878.5
Difficulty Sleeping	Yes	20.8	22.8	46.1%	87.2%	966.2
	No	21.0	23.0	45.9%	87.9%	958.4
	n/a	24.2	24.2	49.9%	96.7%	874.3
Medical Conditions Affecting Thinking	Yes	20.6	22.8	45.6%	86.7%	1014.9
	No	21.0	22.9	46.1%	87.7%	950.8
	n/a	24.2	24.2	49.9%	96.7%	874.5

n/a denotes no response. *Significantly different than respective "Yes" response group (ANOVA; $p \leq 0.01$), but not each other. †Significantly higher response rate than "Yes" and "n/a" response groups (ANOVA; $p = 0.000$). ‡Significantly less (faster) than "Yes" and "n/a" response groups (ANOVA; $p < 0.05$).

across the HealthQScore groups (0–4) are presented in Table 3.

We also differentiated these test scores based on the responses to the individual general health questions (Table 4). The values indicated in Table 4 were calculated considering all valid unique users, regardless of whether they answered the respective question or any of the other general health questions. For nearly every combination of health question and M-CRT performance attribute, users who did not answer the respective health question scored significantly better than those who did. Exceptions to this are noted in Table 4.

The results from our modeling to predict binary (yes/no) classification of the health-related questions and general health status (healthy/unhealthy) based on a calculated HealthQScore are shown in Table 5. Each of these data values in Table 5 indicates the aggregate model performance based on the AUC respective mean derived from the 100 models (10 runs × 10 folds) built for each learner/dataset combination, with the statistically overlapping (confidence interval) highest performing learners for each dataset indicated in bold. Logistic regression was generally the top-performing learner in nearly all cases with moderately robust prediction performance for HealthQScore and the general health questions specific to medications and medical conditions affecting thinking (though, only when using the other three health questions responses as independent variables for the latter).

DISCUSSION

From the original HAPPYneuron program dataset, we cleaned and analyzed individual measures of episodic memory performance from MemTrax and respective selected demographic information from the M-CRT test. Then, using machine learning, we developed a series of models to separately predict the binary classification responses to four individual general health questions and a calculated binary classification index of implied general health status—HealthQScore. Logistic regression

Table 5
Binary classification performance (AUC; 0.0 – 1.0) results for each of the 10 learners

Dataset	5NN	C4.5D	C4.5N	LR	MLP	NB	RF100	RF500	RBF	SVM
MemoryProblems_1Q	0.5489	0.5802	0.5902	**0.5945**	0.5873	0.5841	0.5554	0.5568	0.5850	0.5512
MemoryProblems_4Q	0.5647	0.5863	0.5995	**0.6110**	0.6054	0.5985	0.5704	0.5717	0.5969	0.5722
Medications_1Q	0.6214	0.6532	0.6638	**0.7129**	0.7069	0.7027	0.6480	0.6501	0.6873	**0.7123**
Medications_4Q	0.6962	0.7087	0.7045	**0.7687**	**0.7624**	0.7534	0.7243	0.7261	0.7291	**0.7663**
DifficultySleeping_1Q	0.5270	0.5518	0.5533	**0.5589**	**0.5600**	**0.5636**	0.5286	0.5291	**0.5638**	0.5208
DifficultySleeping_4Q	0.5701	0.5968	0.5989	**0.6247**	**0.6223**	**0.6195**	0.5814	0.5824	0.6133	0.5572
MedicalConditions_1Q	0.5419	0.5025	0.5638	**0.5753**	**0.5758**	**0.5772**	0.5436	0.5451	0.5514	0.5380
MedicalConditions_4Q	0.6767	0.5054	**0.7498**	**0.7532**	**0.7648**	**0.7492**	0.7054	0.7085	**0.7425**	0.6417
HealthQScore_0v4	0.6008	0.5958	0.6162	**0.6802**	0.6599	0.6626	0.5998	0.6028	0.6262	**0.6780**
HealthQScore_01v4	0.5678	0.5237	0.5972	**0.6498**	**0.6392**	**0.6475**	0.5858	0.5873	0.6195	0.5646
HealthQScore_01v34	0.5620	0.6095	0.6049	**0.6475**	0.6312	0.6388	0.5864	0.5886	0.6149	0.6259
HealthQScore_0v34	0.5821	0.6237	0.6261	**0.6800**	0.6510	0.6561	0.6044	0.6053	0.6294	**0.6727**

Statistically overlapping (confidence interval) highest performing learners for each dataset indicated in **bold** (statistically different than all others not in **bold** for the respective model; $p = 0.000$).

was generally the top-performing learning algorithm indicated by its highest or nearly highest AUC performance on all datasets. Classification prediction for HealthQScore was moderately robust, as were the models for the general health questions specific to medications and medical conditions affecting thinking (when the responses to the other three questions were considered as independent variables for the latter). Accordingly, these initial models demonstrate the potential valid clinical utility of MemTrax (administered as part of the M-CRT test) in screening for variations in cognitive brain health. Moreover, we are also introducing supervised machine learning as a modern approach and new value-add complementary tool in cognitive brain health assessment and related patient management.

We created the HealthQScore attribute based on the assumption that a "Yes" response to a greater number of the four M-CRT general health questions suggests a comparatively less overall healthy cognitive state and potentially more likely that the respondent is affected by AD or another form of cognitive impairment. Conversely, users who answered "No" to all the general health questions were assumed to have more likely exhibited normal cognitive brain health at the time of M-CRT participation. Correspondingly, using only a HealthQScore of zero (0) in the negative class resulted in better model performance. Although we currently weighted each of the four general health questions equally in determining a HealthQScore, we recognize that there may be a clinically relevant rationale for weighting these questions differently (singly or in combination) in determining a more appropriate and useful aggregate score.

Nonetheless, there was apparent value in the calculated HealthQScore in differentiating M-CRT performance, in that certain patterns emerged relevant to inferred health status. Whereas selected aspects of M-CRT performance were notably distinct when comparing HealthQScore near extremes (e.g., 0 versus 3 or 1 versus 4), the most consistent progressive pattern of health status differentiation was demonstrated with the true positive response time metric. Moreover, M-CRT performance was also differentiated by the participants' decision to respond to the general health questions, that is, generally those who did not answer a given question (implied to suggest the participant's health was not negatively affected in this respective way) performed better on the M-CRT. This supports our hypothesis that individual health status could be inferred from an aggregate of self-reported indicators and complement (by inclusion) the efficacy of selected features of M-CRT online performance in our preliminary modeling.

Specific to our models targeting individual health questions, it was evident that the models with the other three questions included as independent attributes performed better than those that did not. Without a lot of attributes to consider, adding information from three additional independent attributes potentially makes a larger impact on algorithm learning potential. However, it is also possible that there was some unknown dependency between some of these attributes. For example, including the answers to the other three questions had the greatest effect on the question about medical conditions, raising the highest AUC score by nearly 0.2. It is plausible (though the supporting data are limited) that if someone was taking medications, he or she may have been previously diagnosed with a relevant medical condition. Accordingly, this could be skewing our models. Also, numerous medications prescribed for

a variety of conditions such as anti-cholinergic drugs (including diphenhydramine) and GABA agonists (benzodiazepines, barbiturates, most anti-epileptics) can impair episodic memory and slow reaction time [14–16]. Naturally, our models would likely benefit from, and any underlying dependencies would be clarified by, more definitive questions yielding more precise clinical insight into each individual participant.

Deeper examination of these (or similar) data might prompt select classification algorithm setting changes that would favorably support building more robust models. Interestingly, the models developed for the memory problems question were among the worst performing models for the four general health questions. This was somewhat surprising given that this variation of the dataset contained the most instances, which typically enhances model performance compared to models based on more limited data. Arguably, an underlying reason for this may be that there were still some noisy/faulty data included in the dataset. Further efforts towards additional data cleaning may help improve model performance. Alternatively, while subjective memory complaint can be predictive (in early stages) for future onset and development of dementia, those individuals suffering from or exhibiting cognitive dementia who are diagnosed with AD (beyond mild cognitive impairment) usually deny or are unaware of their memory problems. And, complicating the specificity further, most people recognize and readily admit that their memories are not perfect [17–19].

Clinically, it is especially important and highly valuable to have a simple, reliable, and widely accessible tool to use as an initial screen in detecting early onset cognitive deficits and potential AD. Such *a priori* valid insight would readily reinforce and augment a stratified approach to case management and patient care. Demarcation of relevant functional impairment for research could also be advantageous in stratifying those with early onset cognitive deficits and AD patients in clinical trials to reduce variability and the number of subjects needed and enhance statistical power.

We recognize that this is an early stage in introducing machine learning to cognitive impairment predictive modeling and we realize that the demonstrated model performance in each instance was at best only moderately robust. However, these findings provide a promising indication of how the predictive modeling decision support utility of computerized neuropsychological tests such as MemTrax could be enriched by assessing clinical condition—even if simply via relevant self-reported health questions. Of course, we also recognize that a more definitive clinical diagnosis or assessment of cognitive dysfunction to train the learners would improve predictive model performance and practical clinical utility of MemTrax. Notably, however, a comparison of MemTrax to the recognized and widely utilized Montreal Cognitive Assessment Estimation of mild cognitive impairment underscored the power and potential of this new online tool and approach in evaluating short-term memory in diagnostic support for cognitive screening and assessment with a variety of clinical conditions and impairments including dementia [20]. There is a corresponding urgent need to have quantifiable insight for individuals across the continuum from normal through mild cognitive impairment [7, 21, 22]. A clinically effective MemTrax-based machine learning predictive model could also be instrumental in indicating and tracking the temporal severity and progression of dementia across multiple cognitive and functional domains.

Machine learning has an inherent capacity to reveal meaningful patterns and insights from a large, complex inter-dependent array of clinical determinants and continue to "learn" from ongoing utility of practical predictive models. Thus, we are confident that our models will improve with more and more diverse clinically validated health status data (e.g., a broad multifactorial scope including genomics, promising biomarkers, and other functional, behavioral, and lifestyle indicators) to train the models [2, 11, 23]. A robust, multi-faceted, and externally validated model can uniquely complement and measurably enhance the sensitivity and specificity of MemTrax as a valid cognitive health screen tool and thus greatly assist in clinical decision support and patient management.

Data limitations and outstanding questions

Our initial exploration and assessment of the overall dataset revealed several issues and challenges. Notably, numerous instances of missing information across many features may have compromised the accuracy of our current (and would for any future) models trained on these data. Specifically, the markedly large difference in the number of users who answered whether they were having memory problems compared to the prevalence of responses to the other three general health questions, suggests the need to examine when in the process these questions

were presented to the participants and how the users were prompted.

Whereas our analysis showed significant differences between some features, filter-based modeling (i.e., training models only on a subset of top-ranked features) did not demonstrate meaningful improvement, and thus was not included in the current methods or discussed. The limited number of useful features in these data likely limited the efficacy and utility of this filtering technique that typically is more justified and useful with a greater number of high-value features.

Key predictive modeling findings

- Models for selected health-related questions and the calculated HealthQScore using logistic regression (and several other classifiers) performed moderately well with performance (AUC) in the mid 60 to the mid 70% range.
- These models demonstrate the utility of incorporating MemTrax performance via the M-CRT test in predicting binary health status classification (healthy versus unhealthy) when complemented by select demographic information and only self-reported indirect indicators of general health.
- This novel application of supervised machine learning and predictive modeling helps to demonstrate and validate the cross-sectional utility of MemTrax in assessing early-stage cognitive impairment and general screening for AD.

This illustration is also an important step in advancing the approach for clinically managing this complex condition. By considering and analyzing a wide array of high-value (contributing) attributes across multiple domains of the integrated systems biology and functional behaviors of brain health, informed and strategically directed advanced data mining, supervised machine learning, and robust analytics can be integral (and indeed necessary) to healthcare providers in detecting and anticipating further progression in AD and myriad other aspects of cognitive impairment. Seamless utility and real-time interpretation can notably enhance case management and patient care through innovative technology transfer and commercialization emanating from such models, screening tools, and development of practical and readily usable integrated clinical applications (e.g., via a hand-held device and app). Resulting new insights and discovery will also set the stage for much more significant and impactful future research.

ACKNOWLEDGMENTS

We recognize the work of J. Wesson Ashford, Curtis B. Ashford, and colleagues for developing the on-line continuous recognition task and tool (MemTrax) utilized here and the numerous patients with dementia who contributed to the critical foundational research. We also thank Franck Tarpin-Bernard, his colleagues and assistants, and the participants of the HAPPYneuron program who provided their valuable time and commitment in taking the M-CRT and providing the data for us to evaluate in this study.

J. Wesson Ashford has filed a patent application for the use of the specific continuous recognition paradigm described in this paper for general testing of memory. He also owns the URL for http://www.memtrax.com.

MemTrax, LLC is a company owned by Curtis Ashford, and this company is managing the memory testing system described in this paper and the http://www.memtrax.com URL.

Authors' disclosures available online (https://www.j-alz.com/manuscript-disclosures/19-0165r1).

REFERENCES

[1] Ashford JW, Kolm P, Colliver JA, Bekian C, Hsu LN (1989) Alzheimer patient evaluation and the mini-mental state: Item characteristic curve analysis. *J Gerontol* **44**, P139-P146.

[2] Alzheimer's Association (2016) 2016 Alzheimer's disease facts and figures. *Alzheimers Dement* **12**, 459-509.

[3] Ashford JW, Bernard-Tarpin F, Ashford CB, Ashford MT (2019) A computerized continuous-recognition task for measurement of episodic memory. *J Alzheimers Dis* **69**, 385-399.

[4] Ashford JW, Gere E, Bayley PJ (2011) Measuring memory in large group settings using a continuous recognition test. *J Alzheimers Dis* **27**, 885-895.

[5] Benton AL (1962) The visual retention test as a constructional praxis task. *Confin Neurol* **22**, 141-155.

[6] Buschke H, Fuld PA (1974) Evaluating storage, retention, and retrieval in disordered memory and learning. *Neurology* **24**, 1019-1025.

[7] Ashford JW (2008) Screening for memory disorders, dementia, and Alzheimer's disease. *Aging Health* **4**, 399-432.

[8] Wild K, Howieson D, Webbe F, Seelye A, Kaye J (2008) Status of computerized cognitive testing in aging: A systematic review. *Alzheimers Dement* **4**, 428-437.

[9] Falcone M, Yadav N, Poellabauer C, Flynn P (2013) Using isolated vowel sounds for classification of mild traumatic brain injury. In *2013 IEEE International Conference on*

Acoustics, Speech and Signal Processing, Vancouver, BC, pp. 7577-7581.

[10] Dabek F, Caban JJ (2015) Leveraging big data to model the likelihood of developing psychological conditions after a concussion. *Procedia Comput Sci* **53**, 265-273.

[11] Climent MT, Pardo J, Munoz-Almaraz FJ, Guerrero MD, Moreno L (2018) Decision rree for early detection of cognitive impairment by community pharmacists. *Front Pharmacol* **9**, 1232.

[12] Bergeron MF, Landset S, Maugans TA, Williams VB, Collins CL, Wasserman EB, Khoshgoftaar TM (2019) Machine learning in modeling high school sport concussion symptom resolve. *Med Sci Sports Exerc*. doi: 10.1249/mss.0000000000001903

[13] Van Hulse J, Khoshgoftaar TM, Napolitano A (2007) Experimental perspectives on learning from imbalanced data. In *Proceedings of the 24th International Conference on Machine Learning ACM*, Corvalis, Oregon, USA, pp. 935-942.

[14] Papassotiropoulos A, Gerhards C, Heck A, Ackermann S, Aerni A, Schicktanz N, Auschra B, Demougin P, Mumme E, Elbert T, Ertl V, Gschwind L, Hanser E, Huynh KD, Jessen F, Kolassa IT, Milnik A, Paganetti P, Spalek K, Vogler C, Muhs A, Pfeifer A, de Quervain DJ (2013) Human genome-guided identification of memory-modulating drugs. *Proc Natl Acad Sci U S A* **110**, E4369-E4374.

[15] Pompeia S, Bueno OF, Lucchesi LM, Manzano GM, Galduroz JC, Tufik S (2000) A double-dissociation of behavioural and event-related potential effects of two benzodiazepines with similar potencies. *J Psychopharmacol* **14**, 288-298.

[16] Subhan Z (1984) The effects of benzodiazepines on short-term memory and information processing. *Psychopharmacology Suppl* **1**, 173-181.

[17] Choe YM, Byun MS, Lee JH, Sohn BK, Lee DY, Kim JW (2018) Subjective memory complaint as a useful tool for the early detection of Alzheimer's disease. *Neuropsychiatr Dis Treat* **14**, 2451-2460.

[18] Hill N, Mogle J, Kitko L, Gilmore-Bykovskyi A, Wion R, Kitt-Lewis E, Kolanowski A (2018) Incongruence of subjective memory impairment ratings and the experience of memory problems in older adults without dementia: A mixed methods study. *Aging Ment Health* **22**, 972-979.

[19] Munro CE, Donovan NJ, Amariglio RE, Papp KV, Marshall GA, Rentz DM, Pascual-Leone A, Sperling RA, Locascio JJ, Vannini P (2018) The impact of awareness of and concern about memory performance on the prediction of progression from mild cognitive impairment to Alzheimer disease dementia. *Am J Geriatr Psychiatry* **26**, 896-904.

[20] van der Hoek MD, Nieuwenhuizen A, Keijer J, Ashford JW (2019) The MemTrax Test compared to the Montreal Cognitive Assessment estimation of mild cognitive impairment. *J Alzheimers Dis* **67**, 1045-1054.

[21] Ashford JW, Schmitt FA (2001) Modeling the time-course of Alzheimer dementia. *Curr Psychiatry Rep* **3**, 20-28.

[22] Ashford JW, Shan M, Butler S, Rajasekar A, Schmitt FA (1995) Temporal quantification of Alzheimer's disease severity: 'Time index' model. *Dementia* **6**, 269-280.

[23] Hampel H, O'Bryant SE, Molinuevo JL, Zetterberg H, Masters CL, Lista S, Kiddle SJ, Batrla R, Blennow K (2018) Blood-based biomarkers for Alzheimer disease: Mapping the road to the clinic. *Nat Rev Neurol* **14**, 639-652.