Simple Models for Estimating Dementia Severity Using Machine Learning

W. R. Shankle[ab], Subramani Mani[a], Malcolm B. Dick[c], Michael J. Pazzani[a]
University of California at Irvine, Irvine, California 62967 USA
[a] Dept. of Information and Computer Science, [b] Dept. of Cognitive Science, [c] Dept. of Neurology

Abstract
*Estimating dementia severity using the Clinical Dementia Rating (**CDR**) Scale is a two-stage process that currently is costly and impractical in community settings, and at best has an inter-rater reliability of 80%. Because staging of dementia severity is economically and clinically important, we used Machine Learning (**ML**) algorithms with an Electronic Medical Record (**EMR**) to identify simpler models for estimating total CDR scores. Compared to a gold standard, which required 34 attributes to derive total CDR scores, ML algorithms identified models with as few as seven attributes. The classification accuracy varied with the algorithm used with naïve Bayes giving the highest. (76%) The mildly demented severity class was the only one with significantly reduced accuracy (59%). If one groups the severity classes into normal, very mild-to-mildly demented, and moderate-to-severely demented, then classification accuracies are clinically acceptable (85%). These simple models can be used in community settings where it is currently not possible to estimate dementia severity due to time and cost constraints.*

Key words
Machine Learning, Clinical Dementia Rating Scale, dementia severity, dementia staging, Alzheimer's disease.

Introduction

Dementia due to Alzheimer's disease, and other dementias constitute the fourth most common disorder among the elderly, and has a total cost in the USA of $100 billion annually. Proper treatment can reduce this cost by up to 25%. Early detection of dementia and correct staging of the severity of dementia is critical to selecting the optimal treatment and saving money. We have previously used machine learning (ML) algorithms to improve early detection of dementia by developing a screening test that can be given at home with potentially 94% sensitivity [1].

The objective of the present study is to use ML algorithms to make the most widely used scale for staging dementia clinically practical while preserving accuracy. This scale, the Clinical Dementia Rating Scale [2], in its present mode of implementation, takes 30 minutes and requires trained interviewers, making it unlikely to be widely adopted in general clinical practice. Given an inter-rater reliability of approximately 80% [3,4], it may be possible to eliminate the use of trained interviewers, thereby making the use of the CDR in clinical settings practical. Identifying the key information required to arrive at a total CDR score (rating of dementia severity) and constructing a set of rules that can easily be used to calculate the total CDR score by non-professional personnel is a province for ML research. Previous research on the calculation of the total CDR score has shown that improved rule sets can be achieved [5,6]. However, there has been no research to examine the information required to achieve a reasonably accurate total CDR score. Given the current reliability of 80% using a clinically costly method, we consider the term, "reasonably accurate", to be about 80% or higher. In this first phase of our ML research on the CDR scale, we examine the information needed to compute the total CDR score with an accuracy of 80% or higher. Our focus in this phase is to generate and evaluate simple models to compute the CDR score. These models offer many advantages. For example, they are very tolerant to missing values, as the models encompass only a small number of attributes. By generating competing models, model applicability increases even when there are many missing attributes. Simpler models are also easy to use in a community setting.

In the second phase of this research, we will examine the information needed to compute the six subscale scores of the CDR, then apply an algorithm we have developed to unambiguously compute the total CDR score from these subscale scores. The results of these two approaches will be compared; the one which gives the best combination of clinical practicality and accuracy will be implemented in a high volume dementia clinic that uses medical informatics and ML algorithms to streamline patient care.

Materials and Methods

The EMR of the UCI Dementia Database

The EMR of the UCI Alzheimer's Disease Research Center (**ADRC**) uses a Sybase relational database with a graphical front-end for direct data entry that can be accessed remotely from any platform (MAC, PC, or UNIX). Standardized coding includes the International Classification of Diseases (**ICD9**), and the National Drug Codes (**NDC**). The structure of the medical assessment screens is generic and follows

DeGowin and Degowin's *Bedside Diagnostic Examination* [7]. The database currently holds more than 2,000 patient-visits (patients are longitudinally followed up) and collects more than 1,200 fields per patient-visit. The data used for the present analysis were generated using standard SQL scripts.

The Clinical Dementia Rating Scale of dementia severity

Diagnosing the syndrome of dementia requires the presence of multiple cognitive impairments plus functional impairments resulting from the cognitive impairments in the absence of delirium or other non-organic etiologies such as major depression. The CDR score determines whether a person is demented as well as the severity of the dementia. The total CDR dementia severity score is derived by evaluating the patient for memory, the primary subscale, and five secondary subscales (orientation, judgment and problem solving, community affairs, home and hobbies, and personal care). Each subscale is rated using an ordinal scale (none=0, questionable deficit or very mild dementia=0.5, mild=1, moderate=2, and severe=3). The rater then applies a set of rules to the six subscale scores to obtain the total CDR score, which uses the same severity scale as the subscales.

Development of the gold standard for the total CDR score

Because these rules were confusing to use, even for dementia specialists, we spent two years perfecting an algorithm that computed the total CDR score from the six CDR subscale scores. To do this, a staff neurologist and neuropsychologist independently rated the total and subscale CDR scores for each patient seen at the UCI ADRC. To perfect the computer algorithm for computing total CDR score, we first needed to translate the criteria for assigning CDR subscale scores into a computer program. We did this by selecting attributes from the EMR that measured the subscale categories, then by deriving rules that assigned a score to each of the CDR subscales based on these attribute values. We then compared the computer-derived CDR subscale scores of each patient to those rated by the dementia experts. Differences between expert and computer-derived CDR subscale scores were resolved by group consensus. When it was clear that the CDR subscale score assigned by the program did not apply the subscale scoring criteria as logically as the experts, we modified the program to rectify this problem. However, after rating about 100 patients, the examiners found that, when their CDR subscale scores differed from the computer-derived scores, the algorithm had applied the subscale criteria more logically and appropriately than they had. After examining 302 patients in this fashion, no further modifications of the computer-derived CDR subscale scores were deemed necessary by the professional staff.

In conjunction with this program, the published rules for deriving total CDR score were implemented after eliminating logical errors in those rules [6]. We used a similar process of consensus and revision to rectify any differences between total CDR scores assigned by experts and the algorithm. The programmed algorithms for total and subscale CDR scores were then used to compute the corresponding scores for each patient in the database. If any of the CDR subscale scores could not be computed due to missing attributes, then the total CDR score was not computed. This process produced a gold standard for the total and subscale CDR scores with 100% reliability. To avoid sample bias due to repeated measures, we restricted the sample applied to ML analysis to the first visits of patients in the database; 765 patients had a first visit with sufficient data to compute their total CDR scores.

Attributes used in the ML analysis

The attributes used in the ML analysis consisted of educational level, temporal and spatial orientation from the Mini-Mental Status Exam (MMSE) [8], short term verbal memory from the MMSE and the CERAD DELAYED RECALL TEST [9], short term visual memory from the Wechsler Memory Scale for Visual Reproductions [10], judgment and reasoning from the WAIS-R Information and Similarities subtests [11] and from the clinician's estimation of the patient's insight, plus activities of daily living in the areas of job, hobbies, community activities, household activities, finances, and personal care. These were the same attributes that a computer algorithm used to derive the CDR subscale scores that were then used to derive the gold standard total CDR score as previously described.

Sample Description

The total sample consisted of the initial visits of 765 subjects ranging from normal to severely demented. Table 1 gives their breakdown according to age, sex, education and dementia severity. All subjects were seen at the University of California, Irvine ADRC. Patients received a complete diagnostic evaluation consisting of patient and caregiver interviews, general physical and neurological exam, two hours of cognitive testing including the CERAD [9] neuropsychological battery and other selected tests, routine laboratory testing for memory loss, and magnetic resonance neuroimaging with or without single photon emission with computed tomography. Approximately 50% of the subjects met CERAD criteria for probable or possible Alzheimer's disease [12], 20% met the Alzheimer's Disease Diagnostic and Treatment Center criteria for vascular dementia [13], 7% met diagnostic criteria for Lewy Body dementia, 15% had multiple etiologies, and the remaining 8% were due to a variety of causes. Control subjects were either community volunteers or unaffected spouses of patients, and received an abbreviated, 45 minute version of the patient cognitive battery, which consisted of the CERAD plus measures of activities of daily living. They did not receive a medical exam, laboratory testing or neuroimaging unless cognitive or functional testing suggested an impairment. Dementia severity using the total CDR score was assigned to each subject as previously described.

Table 1 – Sample Characteristics

Attributes	Dementia Staging (CDR total score)			
	Normal (0) n=77	Very Mild (0.5 n=194	Mild (1) n=193	Moderate-Severe (2,3) n=301
Mean Age in years	65.1	69.6	74.2	75.8
% Female	61.8	49.7	56.0	67.4
Mean Years Education	15.3	14.9	13.5	12.3

ML methods

Specific algorithms. We concentrated on decision tree learners (C4.5, CART), rule learners (C4.5Rules) and the naive Bayesian classifier. Decision trees and rules generate clear descriptions of how the ML method arrives at a particular classification. The Naive Bayesian classifier was included for comparison purposes. MLC++ (Machine Learning in C++) [14] is a software package developed at Stanford University which implements commonly used machine learning algorithms. It also provides standardized methods of running experiments using these algorithms. C4.5 [15] is a decision tree generator and C4.5rules produce rules of the form, *if..then* from the decision tree. Naive Bayes [16] is a classifier based on Bayes Rule. Even though it makes the assumption that the attributes are conditionally independent of each other given the class, it is a robust classifier and serves as a good comparison in terms of accuracy for evaluating other algorithms. CART [17] is a classifier which uses a tree-growing algorithm that minimizes the standard error of the classification accuracy based on a particular tree-growing method applied to a series of training subsamples. We used Caruana and Buntine's implementation of CART (the "IND" package), and ran CART twenty times on randomly selected 2/3 training sets and 1/3 testing sets. For each training set, CART built a classification tree where the size of the tree was chosen based on cross-validation accuracy on the training set. The test accuracy of the chosen tree was then evaluated on the unseen test set.

Treatment of missing data. We used each ML algorithm's particular approach for handling missing data. In C4.5 missing attributes are assigned to both branches of the decision node, and the average of the classification accuracy is used for these cases. Therefore, it attempts to learn a set of rules that tolerates missing values in some variables. In the Naive Bayesian Classifier, missing values are ignored in the estimation of probabilities, while CART uses surrogate tests for missing values.

Generation of Training and Testing Samples and analysis. The complete sample was used to randomly assign subjects to either the training or testing set in a 2/3 to 1/3 ratio. This was done 20 times with the complete sample of subjects to generate 20 pairs of training and testing sets. The ML algorithms were trained on the training set and the resulting decision model then classified the unseen testing set. The classification accuracy of each ML algorithm is hence the mean of the accuracies obtained for the 20 runs of the testing set.

Results

Using ML algorithms to estimate the total CDR score from the pool of attributes used to derive the CDR subscale scores gave accuracies ranging from 64%, with C4.5Rules, to 76% with Naïve Bayes (Table 2). These classification accuracies are somewhat misleading because they are due primarily to poor classification of the mildly demented subjects (CDR=1).

Table 2 – Total Accuracy of ML algorithms

Algorithm	Number of runs	Raw accuracy
C4.5	20	68.60
C4.5Rules	20	63.92
naïve Bayes	20	76.47
CART	20	68.37

Table 3 shows the classification accuracies for each CDR dementia severity class based on the results of the Naïve Bayes ML algorithm. All CDR classes have clinically acceptable accuracies except the mildly demented subjects (CDR=1), which are most often misclassified as very mildly demented.

Table 3 – Naive Bayes confusion matrix

True CDR	Estimated CDR Class by Naïve Bayes				% Class Accuracy
	Normal	Very Mild	Mild	Mod-Severe	
Normal	19	5	0	0	79%
Very Mild	0	58	6	2	88%
Mild	0	17	35	8	58%
Mod-Severe	0	4	18	83	79%

To make the ML results more clinically useful, the very mild and mild dementia CDR categories could be grouped in a manner similar to the moderate and severe dementia CDR categories. Such a classification scheme is still useful because the most important clinical distinctions are between normal aging, very mild to mild dementia, and moderate to severe dementia. When grouped in this manner, Table 4 shows that the classification accuracies are quite acceptable. The very mild-to-mild dementia class accuracy is now 92%.

Clinical settings that cannot use a computer algorithm would not be able to use the results obtained from Naïve Bayes

because it does not provide easily understandable decision rules. Furthermore, all the attributes used by Naïve Bayes need to be used to compute the dementia severity class for each patient. In contrast, a tree or rule-based ML algorithm can classify with a subset of the attributes represented in the tree or rule set for individual patients. This feature makes it very clinically attractive. For example, in Figure 1, if a patient's score for orientation to time and place using the items from the MMSE is less than 4.5, no other attributes need to be collected—the patient is classified as moderate to severely demented. We elected to examine the decision tree from CART over C4.5 because, although C4.5 and CART gave similar classification accuracies, the decision tree produced by CART were more compact than C4.5. A neurologist practicing in the community inspected the decision rule sets generated by the 20 CART runs, then ranked them by giving higher priority to those which used inexpensive information at the top of the tree (root) and to those which used the fewest costly tests. Of the twenty decision trees, one tree had clear practical advantages over the rest (Figure 1). It only used two relatively costly attributes (the CERAD 10 item Delayed Free Recall and the WAIS-R Similarities tests) compared to three or more for all the other trees. It also used these attributes only in the subtree pertaining to less demented subjects (those scoring more than 12 points on the orientation questions of the MMSE), such that these more expensive tests could be restricted to higher functioning subjects. Furthermore, the key attributes, mmse orientation to time and place, CERAD Delayed Recall, and WAIS R Similarities, appeared in 17, 19, and 20 of the 20 decision trees generated by CART; mmse orientation and CERAD Delayed Recall were the root node for all but three trees. This means that the generated tree models are reasonably stable.

Table 4 – Naive Bayes confusion matrix recalculated

True CDR	Estimated CDR Class by Naïve Bayes				% Class Accuracy
	Normal	Very Mild-Mild	Mod-Severe	Total	
Normal	19	5	0	24	79%
Very Mild-Mild	0	116	10	126	92%
Mod-Severe	0	22	83	105	79%
Total	19	143	93	255	85%

Discussion

Comparison to the gold standard total CDR dementia severity score Since the ML algorithms did not perform as well as that derived by the human experts, why bother considering the ML algorithms? First, the human experts used 34 attributes to estimate the CDR subscale scores, which were then used with a published algorithm to estimate the total CDR score. Only about 50% of the initial patient visits (765 out of approximately 1,500) had all these attributes. In contrast, the tree selected by the community neurologist required only seven attributes to estimate the total CDR score with accuracy at or above 80%, if one groups the very mild and mildly demented CDR classes. This reduction in number of attributes would considerably increase the proportion of cases for which a total CDR score could be derived. The ML algorithms therefore increase by a wide margin the proportion of cases that can be assessed for dementia severity in community settings, while also reducing the time and expense required for obtaining this information.

Figure 1. The clinically most parsimonious CART model for estimating dementia severity in community practice.

```
mmse time&place orientation < 12.5:
|  mmse time orientation < 4.5:  moderate-to-severe
|  mmse time orientation >= 4.5:
|  |  staying alone < 1.5:
|  |  |  making meals in 2,8,3,1:  mild
|  |  |  making meals in 0:  very mild
|  |  staying alone >= 1.5: moderate-to-severe
mmse time&place orientation >= 12.5:
|  CERAD delayed recall < 5.5:
|  |  CERAD delayed recall < 0.5: moderate-to-severe
|  |  CERAD delayed recall >= 0.5:
|  |  |  writing < 1.5: very mild
|  |  |  writing >= 1.5: mild
|  CERAD delayed recall >= 5.5:
|  |  WAIS R Similarities < 11.5: very mild
|  |  WAIS R Similarities >= 11.5:
|  |  |  housekeeping in 1,8,3,2: very mild
|  |  |  housekeeping in 0:  normal aging
```

Why did the ML algorithms perform at a lower level than an algorithm derived by human experts using a trial-and-error approach with the same attributes? One reason is that the human experts used these attributes to compute the CDR subscale scores, not the total CDR score. Our next set of experiments will be to examine the classification accuracies when ML algorithms use these attributes to estimate the CDR subscale scores directly. Once this is accomplished, the published algorithm for computing the total CDR score can be used. The difference in classification accuracies obtained by the ML algorithms when used directly vs. indirectly (via CDR subscale score estimation), in estimating total CDR score from the same attributes, can then be compared, and the best results implemented in our high volume dementia clinic.

How useful are these results? When the CDR severity categories are grouped into normal, very mild-to-mild, and

moderate-to-severe classes, the classification accuracies parallel that obtained from human experts using an extensive interviewing process. For many clinical practice settings, this coarser classification scheme is adequate, and still makes the important distinctions between normal and demented, and between mild and more-than-mild dementia. Without the data reduction provided by ML algorithms, it is likely that community clinicians would estimate dementia severity (if they do it at all) much worse than the 80% inter–rater reliability obtained by experts. The ML algorithms therefore provide a substantive contribution to the practical estimation of dementia severity in community settings.

It is clear from our and others work that ML algorithms can optimize clinical practice guidelines and streamline costs of health care delivery. However, the ability for ML algorithms to do this is frequently constrained by an insufficient number of exemplars, particularly when samples are subdivided to address questions more precisely. This problem emphasizes the importance of developing facile, structured data collection methods for clinical medicine so that the true potential of Machine Learning for intelligent decision support is not held back by insufficient data.

Conclusion

This initial set of experiments in estimating the Clinical Dementia Rating severity score shows that Machine Learning algorithms achieved a substantial reduction in information and cost of information with only a minimal reduction in accuracy. The ML methods generated very simple models for assessing dementia severity which could be employed in community settings with considerable ease. Further experiments will be directed at estimating the subscale scores of the CDR to determine whether accuracy can be further improved.

References

[1] Shankle, WR., Mani, S., Pazzani, M., and Smyth, P. Detecting very early stages of dementia from normal aging with machine learning methods. In Keravnou, E., Garbay, C., Baud, R., and Wyatt, J., editors, Lecture Notes in Artificial Intelligence: Artificial Intelligence in Medicine, AIME97, volume 1211, pages 73–85. Springer, 1997.

[2] Hughes, CP., Berg, L., Danziger, WL., Coben, LA and Martin, RL. A new clinical scale for the staging of dementia. British Journal of Psychiatry, 140:566–72, June 1982.

[3] Burke, WJ., Miller, JP., Rubin, EH., Morris, JC., Coben, LA., Duchek, J., Wittels, IG., Berg, L. Reliability of the Washington University Clinical Dementia Rating. Archives of Neurology, 45(1):31–2, 1988.

[4] McCulla, MM., Coats, M., Van Fleet, N., Duchek, J., Grant, E., and Morris, JC. Reliability of clinical nurse specialists in the staging of dementia. Archives of Neurology, 46(11):1210–1, Nov, 1989.

[5] Gelb, DJ and St. Laurent, RT. Alternative calculation of the global clinical dementia rating. Alzheimer Disease and Associated Disorders, 7(4):202–11, 1993.

[6] Alzheimer's Disease Cooperative Study Unit (Manual). Assignment of Clinical Dementia Rating. Jan, 1994.

[7] E.L. DeGowin and R.L. DeGowin. Bedside Diagnostic Examination. Macmillan, New York, 7th edition, 1976.

[8] Folstein, MF., Folstein, SE., and McHugh, PR. Mini-mental state—A practical method for grading the cognitive state of patients for the clinician. Journal of Psychiatric Research, 12(3):189–98, Nov, 1975.

[9] Welsh, KA., Butters, N., Mohs, RC., Beekly, D., Edland, S., and Fillenbaum, G. The Consortium to Establish a Registry for Alzheimer's Disease (CERAD) Part V—A normative study of the neuropsychological battery. Neurology, 44(4):609–14, Apr, 1994.

[10] Prigatano, George P. Wechsler memory scale: a selective review of the literature. Clinical Psychology, 1978. Series title: Archives of the behavioral sciences monograph series; no. 54.

[11] Wechsler, David. Manual for the Wechsler adult intelligence scale. Psychological Corp, New York, 1955.

[12] G McKhann, D Drachman, M Folstein, R Katzman, D Price, and EM Stadlan. Clinical diagnosis of Alzheimer's disease: Report of the NINCDS-ADRDA work group under the auspices of the department of health and human services task force on alzheimer's disease. Neurology, 34(7):939–44, Jul 1984.

[13] Chui, J.I. Victoroff, D. Margolin, W. Jagust, R. Shankle, and R. Katzman. Criteria for the diagnosis of ischemic vascular dementia proposed by the state of California Alzheimer's disease diagnostic and treatment centers. Neurology, 42(3):473–80, Mar 1992.

[14] R Kohavi, George John, Richard Long, David Manley, and Karl Pfleger. MLC++: A machine learning library in C++. In Tools with Artificial Intelligence, pages 740–743. IEEE Computer Society Press, 1994.

[15] JR Quinlan. C4.5: Programs for Machine Learning. Morgan Kaufmann, Los Altos, California, 1993.

[16] RO Duda and PE Hart. Pattern Classification and Scene Analysis. John Wiley, New York, 1973.

[17] L Brieman, J.H. Friedman, R.A. Olshen, and C.J. Stone. Classification and Regression Trees. Wadsworth, Belmont, 1984.

Address for correspondence

Subramani Mani, Dept. of Information and Computer Science, University of California at Irvine, Irvine CA 62967 USA. Ph: 714-824-1316 Fax: 714-824-4056
mani@ics.uci.edu http://www.ics.uci.edu/~mani