Original Paper

Neuro
epidemiology

Neuroepidemiology 2010;35:19–27
DOI: 10.1159/000302844

Received: August 13, 2009
Accepted: December 19, 2009
Published online: April 2, 2010

Analyses of Diagnostic Patterns at 30 Alzheimer's Disease Centers in the US

Kyle Steenland[a, b] Jessica Macneil[a] Scott Bartell[c] James Lah[b]

[a]Rollins School of Public Health and [b]Alzheimer's Disease Research Center, Emory University, Atlanta, Ga., and [c]Program in Public Health, University of California at Irvine, Irvine, Calif., USA

Key Words
Alzheimer's disease · Mild cognitive impairment · Cognition

Abstract
Background: The US Alzheimer's Disease Centers (ADCs) (n = 30) recently created a uniform data set. We sought to determine which variables were most important in making a diagnosis, and how these differed across ADCs. ***Methods:*** A cross-sectional analysis of first visits to ADCs via polytomous logistic regression. We analyzed subjects with complete data (n = 7,555, 89%), and also used multiple imputation to infer missing data. ***Results:*** There were 8,495 subjects; 50, 26, and 24% were diagnosed as normal, having mild cognitive impairment (MCI), or mild Alzheimer's disease [Clinical Dementia Rating (CDR) score <1], respectively. The model using 7,555 subjects was 86% accurate in predicting diagnosis. Important predictors were physician-reported decline and the CDR sum of boxes, followed by 4 cognitive tests (Mini Mental State Examination, Category Fluency Tests, Logical Memory Test, Boston Naming Test). Multiple imputation revealed Trail Making Test B to be additionally important. Consensus versus single-clinician diagnoses were 2–3 times more likely to result in MCI than normal diagnoses. Excluding clinical judgment variables, functional assessment and psychiatric symptoms were important additional predictors; model accuracy remained high (78%). There were significant differences between centers in the use of different cognitive tests in making diagnoses. ***Conclusions:*** We recommend creating a hypothetic data set to use across ADCs to improve diagnostic consistency, and a survey on the use of raw or adjusted cognitive test scores by different ADCs.

Introduction

The National Alzheimer's Coordinating Center (NACC) initiated the collection of uniform data (uniform data set; UDS) from approximately 30 US Alzheimer's Disease Centers (ADCs) in 2005 [1, 2]. The UDS covers demographics, cognitive tests, assessment of functional abilities, medical history, family history, clinical impressions, and diagnoses. As of August 2007, the UDS included data from approximately 10,000 initial visits and a smaller number of follow-up visits.

It is known that among clinicians there is considerable heterogeneity in diagnosing mild cognitive impairment (MCI) and mild dementia [3, 4]. We have analyzed the UDS data to (1) assess differences in diagnostic patterns across the ADCs, and (2) construct a best possible model of key variables to predict observed diagnoses.

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

Accessible online at:
www.karger.com/ned

Dr. Kyle Steenland
Emory University, Rollins School of Public Health
1518 Clifton Road
Atlanta, GA 30322 (USA)
Tel. +1 404 712 8277, E-Mail nsteenl@sph.emory.edu

Methods

NACC UDS data were available for approximately 2 years, from start-up in August 2005 through August 2007. We restricted our analyses to subjects with a global Clinical Dementia Rating (CDR) score of ≤1, who were classified as normal, having MCI, or mild probable Alzheimer's disease (AD) (n = 8,495). Subjects with cognitive impairment but not MCI were excluded.

In building our predictive models, we initially included all UDS variables that we considered possibly relevant. These were age, sex, race, education, first-degree relatives with dementia, 4 medical comorbidities (high blood pressure, heart disease, cerebrovascular disease, and diabetes), clinician-determined current depression, Geriatric Depression Scale, clinically judged alcohol abuse, smoking, body mass index, Hachinski score, decline reported by physician (in either memory, other cognitive domain, function, or behavior), time since decline began, global CDR, CDR sum of boxes, Neuropsychiatric Inventory Questionnaire (NPI-Q) summary score, NPI-Q severity scores, Functional Activities Questionnaire (FAQ), 9 cognitive tests [Mini Mental State Examination (MMSE), Logical Memory Test, Category Fluency Tests (sum of animals and vegetables), Boston Naming Test, Digit Span Forward Test (total trials), Digit Span Backward Test (total trials), Trail Making Tests A and B, and Digit Symbol Test], and whether the diagnosis was made by consensus or by a single clinician. A recent review article has described many of these tests and recent results for them in a clinically normal population [5]. See Appendix 1 for a brief description of neurobehavioral tests and clinical assessments.

A number of these variables were highly correlated with each other (e.g. different versions of the same test) and we tried different combinations of correlated variables in reduced models including only age as a covariate, to determine which of two correlated variables was a better predictor of diagnosis. For example, CDR sum of boxes was a better predictor than CDR global score. NPI-Q severity scores were not as predictive as yes/no for NPI-Q items, and the sum of 5 NPI-Q items (anxiety, apathy, disinhibition, irritability, night behaviors) performed better than the sum of all 12 NPI-Q items. We defined NPI-Q sum as a dichotomous variable if the sum of these 5 items was 1 or more, and 0 otherwise. We used immediate logical memory instead of delayed logical memory; these 2 tests were essentially equivalent, with a Spearman correlation coefficient of 0.91. Ultimately, we were left with 29 variables which were potentially predictive of diagnoses and candidates for inclusion in predictive models. Table 1 lists these variables as well as the percent of subjects missing data for each of them.

The cognitive test values in the UDS are not adjusted for demographic variables such as age, sex, and education. It is likely that some centers used scores for these tests which are normalized for these demographic variables, which could contribute to variability across centers in using these tests.

Our general approach was to construct a polytomous logistic model to predict all three outcomes (normal, MCI, AD) via the same model, using SAS PROC LOGIST with a g-logit link (SAS version 9.1, SAS Institute, Cary, N.C., USA). This model estimates simultaneously the odds ratio (OR) for a diagnosis of MCI versus normal, and for a diagnosis of AD versus MCI. ORs are not constrained to be proportional.

We adopted two different approaches to building a logistic model to predict diagnoses, corresponding to goals 1 and 2 mentioned above.

Available Case Approach

For our first goal, assessing heterogeneity across centers, we sought a parsimonious model of important variables in which we could test interaction terms between center and other variables. In this analysis, we sought to maximize the number of subjects included, but yet ensure that all included subjects had no missing data.

We first built a logistic model using 5,474 subjects (64% of all 8,495 subjects) with complete data on all 29 variables which we thought were potentially predictive of diagnosis (table 1). Using backward selection we restricted the model to variables which were statistically significant predictors at the p = 0.05 level (for the combined effect for the 2 ORs, i.e. a 'chunk test' for the parameters for the OR for MCI vs. normal, and AD vs. MCI). This resulted in a model with 17 variables (those in table 3, plus Trail Making Test A, Digit Span Forward Test, Digit Symbol Test of the Wechsler Adult Intelligence Scale, and Trail Making Test B). Two of these variables were missing for a large number of subjects: the cognitive tests Trail Making Test B and Digit Symbol Test were missing for 10 and 12% of all subjects, respectively. As our goal was to build a parsimonious model with as many subjects as possible for testing heterogeneity across centers, we excluded these two variables from the model, and re-ran the backward selection procedure, which resulted in inclusion of 13 variables (table 3) in the model (Trail Making Test A and Digit Span Forward Test were no longer significant). Running the model with these 13 variables allowed inclusion of 7,555 subjects (89% of all subjects), all of whom had complete data for all 13 variables.

The final model using only complete data included variables in table 3. Education and age were treated as continuous variables, while depression, presence of a first-degree relative with dementia, and race (white/non-Hispanic vs. all other) were treated as dichotomous variables. All cognitive tests and clinical variables were treated as continuous variables, with the exception of physician-reported decline which was dichotomous. Approximate linearity in exposure-response was checked for continuous variables via categorical analyses prior to treating them as continuous.

We assessed differences between centers by including interaction terms between each center and different demographic and cognitive test variables. We did not test interactions between center and Hachinski score, clinician-reported decline, CDR sum of boxes, and consensus diagnosis, because models generally did not converge due to lack of variation in data for some outcomes in some centers. Separate interaction models were run for each demographic or cognitive variable of interest. For example, 30 models (1 for each center) were run for each demographic variable (and each cognitive test variable), each of which included a single interaction term between a specific center and that demographic variable (or cognitive variable). A significant interaction (at the p = 0.05 level) between center X and age would indicate that center X gave a different importance to age than other centers in making their diagnosis, adjusted for other demographic variables, clinical variables, and cognitive test variables. As a summary, we present the number of significant interaction terms across the 30 centers, with 1–2 significant terms expected by chance.

Table 1. Candidate variables from the UDS

Variable	Available	Missing	Missing %	Normal missing, %	MCI missing, %	AD missing, %
Physician-reported decline	8,495	0	0.0	0.0	0.0	0.0
Depression	8,495	0	0.0	0.0	0.0	0.0
Gender	8,495	0	0.0	0.0	0.0	0.0
Consensus diagnosis	8,495	0	0.0	0.0	0.0	0.0
Heart disease	8,495	0	0.0	0.0	0.0	0.0
CDR sum of boxes	8,495	0	0.0	0.0	0.0	0.0
Age	8,495	0	0.0	0.0	0.0	0.0
Cerebrovascular disease	8,495	4	0.0	0	50	50
Diabetes	8,495	22	0.3	55	36	9
High blood pressure	8,495	38	0.4	50	32	18
Race	8,442	53	0.6	45	27	28
Years of education	8,430	65	0.8	63	17	20
MMSE	8,410	85	1.0	34	37	29
First-degree relative demented	8,382	113	1.3	43	28	29
Alcohol abuse	8,326	169	2.0	44	31	25
Tobacco use (100 cigarettes)	8,325	170	2.0	44	31	25
Hachinski Ischemia Score	8,256	239	2.8	77	16	7
Functional Assessment Questionnaire	8,252	243	2.9	59	27	14
Geriatric Depression Scale	8,228	267	3.1	36	24	40
Digit Span Forward Test	8,201	294	3.5	34	21	45
Digit Span Backward Test	8,188	307	3.6	32	21	47
Category Fluency Tests	8,151	344	4.0	34	22	44
Trail Making Test A	8,133	362	4.3	29	19	52
Boston Naming Test	8,121	374	4.4	34	23	43
Logical Memory Test	8,095	400	4.7	34	23	43
NPI-Q category	7,848	647	7.6	62	22	16
Trail Making Test B	7,673	822	9.7	17	15	68
Wechsler Adult Intelligence Scale (Digit Symbol Test)	7,514	981	11.5	41	21	38
Body mass index	8,432	995	11.7	62	32	26

5,474 subjects (64%) had complete data on all variables. See Appendix 1 for a description of cognitive and clinical tests.

Table 2. Mean (SD) of cognitive tests for three diagnostic groups in the UDS

Test	Normal	MCI	AD (CDR ≤1)
MMSE (range 1–30, higher is better)[1]	28.9 (1.4)	27.0 (2.7)	21.7 (4.8)
Logical Memory Test (range 0–25, higher is better)	13.7 (4.0)	9.4 (4.4)	4.3 (3.5)
Category Fluency Tests (range 0–98, higher is better)	34.1 (8.7)	26.4 (7.6)	18.1 (7.5)
Boston Naming Test (range 0–30, higher is better)	27.0 (3.4)	24.3 (5.0)	19.8 (6.9)
WAIS (Digit Symbol Test) (range 0–93, higher is better)	46.0 (12.9)	36.0 (12.8)	26.1 (13.2)
Trail Making Test B (range 10–300, higher is worse)	94.6 (54.1)	146.6 (79.5)	209.9 (86.7)
Trail Making Test A (range 1–150, higher is worse)	35.9 (17.2)	48.1 (25.7)	69.2 (39.8)
Digit Span Backward Test (range 0–12, higher is better)	6.8 (2.2)	5.7 (2.1)	4.7 (1.9)
Digit Span Forward Test (range 0–12, higher is better)	8.5 (2.1)	7.7 (2.1)	7.1 (2.2)
FAQ (range 1–30, higher is worse)	0.5 (1.9)	3.7 (5.1)	13.8 (7.7)

WAIS = Wechsler Adult Intelligence Scale. Ten and twelve percent of subjects missing data for Trail Making Test B and WAIS, respectively. [1] n = 8,495. Range reported is observed range.

Table 3. ORs for AD versus MCI, and MCI versus normal: model using complete data

Variable	MCI versus normal			AD versus MCI		
	OR	χ^2	p value	OR	χ^2	p value
Clinician-reported decline versus none	20.12	592.8	<0.0001	4.72	18.4	<0.0001
CDR sum of boxes (per unit of CDR sum, higher is worse)	2.33	88.3	<0.0001	2.65	517.2	<0.0001
MMSE (per unit increase)	0.83	39.6	<0.0001	0.84	69.2	<0.0001
Consensus versus single-clinician diagnosis	3.12	69.6	<0.0001	0.54	23.4	<0.0001
Category Fluency Tests (per unit increase)	0.95	65.2	<0.0001	0.96	20.9	<0.0001
Logical Memory Test (per unit increase)	0.93	39.5	<0.0001	0.93	20.9	<0.0001
Boston Naming Test (per unit increase)	0.92	34.0	<0.0001	0.96	13.6	0.0002
Hachinski Ischemia Score (per unit increase, higher suggests vascular dementia)	1.00	0.0	0.94	0.78	35.8	<0.0001
Education (OR per year of schooling)	1.06	13.3	0.0003	1.05	9.1	0.0025
Race (white vs. black/Hispanic)	1.03	0.1	0.82	1.85	15.7	<0.0001
First-degree relative demented versus no such relative	1.12	1.4	0.23	1.51	13.2	0.0003
Depression versus no depression	0.69	6.3	0.01	0.66	7.6	0.005
Age (OR per 5 years of age)	0.96	2.6	0.11	0.93	5.0	0.03

7,555 people included in the model who had data on all variables. A χ^2 value greater than 3.84 indicates that a parameter was statistically significant at the 0.05 level. Variables sorted in order of importance, based on sum of 2 χ^2 values.

To assess the predictive accuracy of the complete data model, each subject was assigned a predicted outcome based on the highest of 3 outcome probabilities resulting from the model, and the predicted outcome was compared with the observed outcome. Accuracy was calculated as the percent of subjects for whom diagnosis (either normal, MCI, or AD) was correctly predicted. Because accuracy using results for all subjects can be inflated, we also performed a 10-fold cross-validation [6–8] to calculate accuracy. In this technique, 10% of the observations are sequentially withheld and a model fit to the remaining 90%; this process is repeated 10 times, resulting in 10 different sets of estimated effects for the 13 variables in the model. Each set of effect estimates was then used to predict diagnoses for the corresponding 10% of excluded observations. Finally, predicted diagnoses were compared with observed diagnoses, across all of the 10% excluded data sets, to obtain the 10-fold cross-validation estimate of accuracy.

Imputation Approach

Our second approach corresponded to our goal of building a model to accurately predict diagnosis, using all relevant variables in table 1, while addressing the problem of missing data. We used multiple imputation techniques to estimate any missing values among all 8,495 subjects, and then built a model using backward selection (p = 0.05 to retain). This approach resulted in a model with 18 variables and was based on data from all 8,495 subjects.

This imputation approach was motivated because analyses restricted to subjects with complete data may be biased if missing values are not missing completely at random (MCAR) [9], i.e., if 'missingness' depends on any other variables. In our data, the MCAR assumption was unlikely to hold. For example, those missing data for the Digit Symbol Test and Trail Making Test B were more likely to have had AD than to have been diagnosed as either normal or having MCI.

Multiple imputation assumes that the probability of missingness is only dependent on observed values, a less restrictive assumption than MCAR [9]. Hence data for missing variables for any given subject can be imputed based on their values for variables which were observed. The imputation is done by fitting models to predict missing values for a given variable (e.g. variable A) based on observed values of all other variables, including the outcome variable. In addition, the error term from the imputation model for variable A is used to assign randomly (based on a normal distribution) an error component to the predicted (imputed) value for variable A.

First, 50 data sets were generated with imputation of missing values using default predictive mean matching and logistic regression imputation via the 'mice' library in R [10, 11]. These 50 data sets were used to build a polytomous logistic regression model, with variables selected via backward selection (as in the complete data approach). At each step of the backwards model selection procedure, one polytomous regression model was fit to each of all 50 imputed data sets; parameter estimates were obtained by averaging over all 50 sets of parameter estimates, and variance estimates were obtained via the usual combination of between- and within-imputation variances. Wald test scores for each predictor were calculated from these multiple-imputation-derived parameter estimates, and the least significant predictor was dropped from the next step. These Wald tests were 'chunk' tests for 2 parameters at a time for each variable, i.e., the OR for MCI versus normal, and for AD versus normal (a Wald test with 2 d.f.).

To examine more thoroughly the role of variables other than those obtained by clinical judgment, we also used imputation to run a backwards selection model after 4 variables based on clinical judgment (clinician-reported decline, CDR sum of boxes, clinician-reported depression, and consensus diagnosis were removed from the list of candidate variables in table 1).

Once the backwards selection procedure was completed, we again calculated the accuracy of the model using the 10-fold cross-validation as was described above for the complete data analysis. This technique was applied for the final model using each of the 50 imputed data sets, and results averaged across all 50 data sets.

Results

There were 4,241 normal subjects (50%), 2,198 MCI subjects (26%), and 2,056 AD subjects (24%) with initial visits to 30 ADCs. The number of subjects per center varied from 75 to 510 with a mean of 283. Forty-one percent of subjects were male, 79% white, non-Hispanic, 29% had a high school education or less, 13% were judged depressed, 48% reported a first-degree relative with dementia, and the average age was 75. Descriptive statistics on 9 cognitive tests and the FAQ are shown in table 2.

Available Case Approach and Tests for Heterogeneity between Centers

As noted, the 'available case' analysis used 7,555 subjects (89% of the 8,495 subjects) with complete data on 13 predictors. Results are shown in table 3. Variables in this table are sorted by order of importance, based on the sum of the χ^2 tests of significance for the 2 ORs (MCI vs. normal, AD vs. MCI). The most important predictors were physician-reported decline ('does the physician believe there has been a current meaningful decline in the subject's memory, nonmemory cognitive abilities, behavior, ability to manage his/her own affairs, or have there been motor/movement changes?'), the CDR sum of boxes, the cognitive tests, and whether the diagnosis was made by consensus or not.

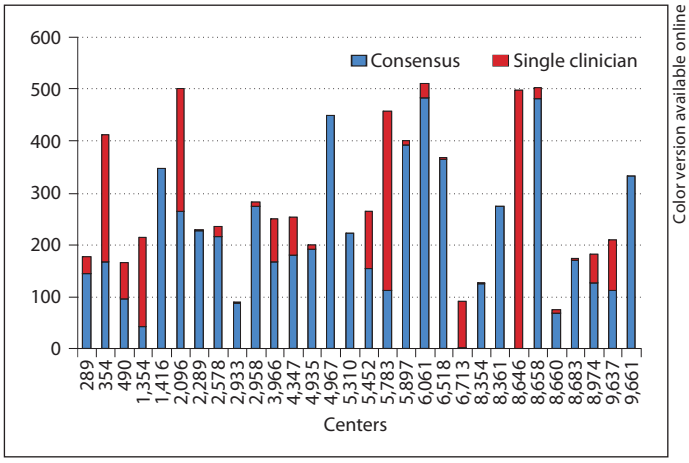
The ORs for many variables have an intuitive interpretation. For example, given the same demographic characteristics, a subject with higher MMSE scores is less likely to be diagnosed as having MCI versus normal, and less likely to be diagnosed as having AD versus MCI (ORs less than 1.0). For example, each point increase in the MMSE decreases the odds of a diagnosis of MCI versus normal by 14%. Results for CDR sum of boxes indicate that for each unit increase, the odds of being diagnosed as having AD versus MCI increases 2.6-fold, and the odds of being diagnosed as having MCI versus normal increases 2.4-fold.

The interpretation of the ORs for demographic variables can be counterintuitive. For example, given the same set of cognitive test scores, a higher age makes it less likely (OR $<$ 1) for a subject to be diagnosed as having MCI versus normal, and less likely to be diagnosed as demented versus MCI. On the other hand, higher education has the opposite effect (OR $>$ 1), increasing the likelihood of a worse diagnosis. Being of white, non-Hispanic race/ethnicity also increases the risk of a worse diagnosis (OR $>$ 1), given the same set of cognitive test scores. In all these cases, it is important to appreciate the difference of these variables as risk factors for worse outcome in the 'world at large' versus their role in determining diagnosis within an ADC given a set of cognitive and clinical variables for a patient. Age increases the risk of worse cognition and hence the risk of incident disease in the 'world at large'. Yet given a set of cognitive scores and clinical variables, within an ADC those with higher age are more likely to get a better diagnosis. This may be partly a reflection of the availability of only raw data for age in the UDS, rather than standardized z-score. It may also simply reflect the routine clinical judgment that for a given set of cognitive scores, raw or age-standardized, older patients are judged to have a less severe diagnosis than younger patients.

Overall, the logistic model in table 3 is a good predictor of the 3 outcomes. The model accurately classified 86% of the subjects, with or without cross-validation.

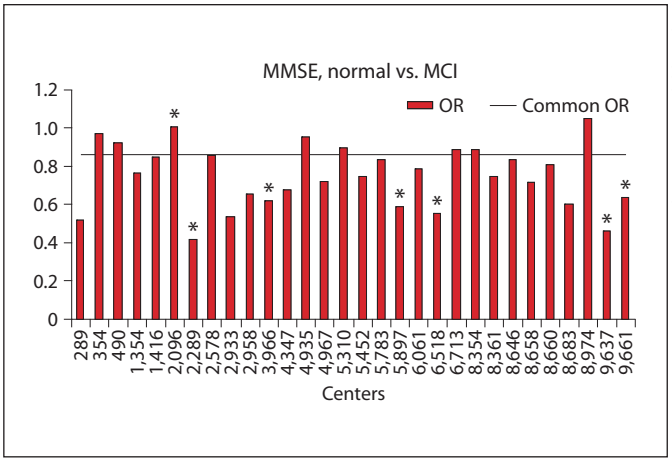
Seventy-four percent of diagnoses across all centers were made by consensus versus a single clinician (fig. 1). The OR for diagnosing MCI versus normal based on consensus diagnoses compared to single-clinician diagnoses was 3.12 (95% CI = 2.48–3.92). In contrast, diagnoses of AD based on consensus were half as likely compared to diagnoses of MCI (OR = 0.54; 95% CI = 0.42–0.69). The importance of consensus diagnosis, however, may be confounded by ADC, since 15 centers have less than 5% of their diagnoses made by single clinicians (4 have none), and 2 centers have only single-clinician diagnoses, making it difficult to separate out the effects of center versus consensus diagnosis. Limiting the data to 13 centers with at least 5% of diagnoses made by single clinicians, and excluding 2 centers with only single-clinician diagnoses (leaving 43% of the overall data), the OR for a consensus versus a single-clinician diagnosis resulting in MCI rather than normal remains markedly elevated (OR = 2.93, 95% CI = 1.60–3.74), but the OR for a consensus versus a single-clinician diagnosis resulting in AD versus MCI now is much less marked (OR = 0.90, 95% CI = 0.61–1.33) and no longer statistically significant (p = 0.53).

Variation between centers in the weight they gave to specific variables was substantially greater than would be expected by chance. Figure 2 shows the center-specific ORs for diagnosing MCI versus normal, per unit change



Color version available online

Fig. 1. Consensus diagnoses by center.



Color version available online

Fig. 2. ORs for diagnosing MCI versus normal, per unit increase in MMSE (significant differences from common OR indicated by an asterisk).

in the MMSE. The common OR across all centers was 0.86 for MCI versus normal, for each 1-point increase in the MMSE. Seven centers differed significantly from the common ORs for MCI versus normal (vs. 1–2 significant differences predicted by chance).

Table 4 summarizes the number of significant interaction terms between center and specific demographic variables or cognitive tests. There was more variation between centers than would be expected by chance; the cognitive tests showed more intercenter variation than the demographic variables.

Multiple Imputation Approach

Table 5 gives the results using the multiple imputation approach which included data for all 8,495 subjects. The general pattern of results is similar to the complete data approach in table 3. Clinician-reported decline and CDR sum of boxes are the most important predictors, consensus diagnosis remains important, and the most important cognitive tests remain MMSE, Category Fluency Tests, Logical Memory Test, and Boston Naming Test. However, Trail Making Test B, omitted in the complete data approach due to a high percentage of missing data, is seen to be an important cognitive test. To a lesser extent, this is also true for the Digit Symbol Test. The accuracy of the model in table 5 is again 86% (using 10% cross-validation).

We then re-ran the backward selection procedure after excluding 4 variables based on clinical judgment (clinician-reported decline, CDR sum of boxes, physician-reported depression, and whether the diagnosis was based on consensus or not) from candidate variables in table 1. Results (not shown) were similar to table 5, with some notable differences. Without variables based on clinical judgment, the FAQ becomes the most important predictor. FAQ is highly correlated with CDR sum of boxes and

Table 4. Number of centers differing significantly from common OR, demographic variables

	Significant[1] interactions center × variable, for AD vs. MCI	Significant interactions center × variable, MCI vs. normal	Significant interactions expected by chance
Demographic variables			
Age	3	5	1–2
Race = white, non-Hispanic	2	4	1–2
Education	4	3	1–2
Depression	2	4	1–2
Relative demented	0	5	1–2
Cognitive tests			
MMSE	5	7	1–2
Logical Memory Test	4	6	1–2
Category Fluency Test	8	5	1–2
Boston Naming Test	6	6	1–2

[1] Interaction significant at $p < 0.05$ level, 30 × 5 models run for demographic interactions (30 centers and 5 demographic variables), and 30 × 4 models run for cognitive test interactions (30 centers, 4 cognitive tests). Each model included the variables in table 2, and a variable for center, and either an interaction term between a demographic variable and center, or an interaction term between a cognitive test and center.

Table 5. ORs for AD versus MCI, and MCI versus normal: model using imputed data

	MCI versus normal			AD versus MCI		
	OR	χ^2	p value	OR	χ^2	p value
Clinician-reported decline	20.45	654.1	<0.0001	4.26	19.9	<0.0001
CDR sum of boxes (per unit)	2.29	79.5	<0.0001	2.42	324.0	<0.0001
Consensus versus single-clinician diagnosis	3.24	108.1	<0.0001	0.53	26.8	<0.0001
MMSE (per unit)	0.86	31.0	<0.0001	0.85	64.8	<0.0001
Logical memory delayed (per unit)	0.92	43.3	<0.0001	0.93	23.7	<0.0001
Category Fluency Test (per unit)	0.96	41.3	<0.0001	0.97	14.3	0.0002
Education (per year of schooling)	1.09	31.6	<0.0001	1.07	15.5	0.0001
Hachinski Ischemia Score (per unit)	0.98	0.3	0.59	0.77	42.3	<0.0001
Boston Naming Test (per unit)	0.94	25.8	<0.0001	0.95	15.3	0.0001
Trail Making Test B (per 10 units)	1.05	25.0	<0.0001	1.04	16.0	0.0001
Age (per 5 years)	0.90	17.6	<0.0001	0.91	10.0	0.002
Race (white vs. black/Hispanic)	1.19	2.0	0.16	2.09	22.8	<0.0001
Digit Symbol Test (per 10 units)	0.79	21.2	>0.0001	1.00	0	1.0
First-degree relative demented	1.16	2.4	0.12	1.56	16.4	0.0001
Trail Making Test A (per 10 units)	1.11	11.1	0.0009	1.04	4.0	0.05
Depression	0.62	9.0	0.003	0.82	1.7	0.20
FAQ (per 5 units)	1.03	0.1	0.78	1.18	9.0	0.003
Geriatric Depression Scale (per 5 units) (higher indicates more depression)	1.04	0.1	0.71	0.74	7.0	0.008

8,495 people included in the model. A χ^2 value greater than 3.84 indicates that a parameter was statistically significant at the 0.05 level. Variables sorted in order of importance based on sum of 2 χ^2 values.

physician-reported decline (Spearman correlations, 0.85 and 0.73, respectively), which presumably explains why it plays an important predictive role, when these 2 clinical variables are not included in the model. To a lesser extent, the same is true for NPI-Q score (correlations, 0.49 and 0.41 with CDR sum of boxes and physician-reported decline, respectively). The Digit Symbol Test is also added to this model, while depression drops out. The accuracy of this model without clinical variables is 78% (using 10% cross-validation).

Discussion

We have built several predictive models to predict diagnosis of normal, MCI, and mild AD across all 30 ADCs that have contributed UDS data to the NACC. Our predictive models, whether based on complete data with 88% of subjects, or based on all subjects with imputation of missing data, accurately predicted the final diagnosis in 86% of cases, using cross-validation techniques.

Many of the cognitive tests in the UDS are known to vary by age, race, education, and sex among subjects judged clinically normal [5]. Our data suggest that younger age, higher education, and white race – given the same cognitive test values – are associated with a worse diagnosis. This finding may result from some clinicians using norms for age, race, and education, resulting in age-, race- and education-adjusted cognitive scores, while we have access only to raw scores from the cognitive tests. If norm-adjusted cognitive tests were used commonly across centers, while we have only raw scores, we would see demographic effects in our analyses, which might disappear if we were to have adjusted scores for cognitive tests.

The availability of norms varies widely by cognitive test. While norms for age exist for all 9 cognitive tests considered in our data, norms for race exist only for 4, and norms for education exist for only 5. Therefore, it would not be surprising that race and education show demographic effects in our data, given that for many cognitive tests in our model, no adjustment has been made for race or education. On the other hand, the fact that our data show the effect of age is surprising given that presumably most clinicians in the ADCs used age-adjusted data for all cognitive tests. However, anecdotal informa-

tion suggests that even when using age-adjusted scores, those of younger age with the same cognitive test scores, an older person may be judged clinically to have a worse diagnosis.

On the other hand, our results of the relative importance of the cognitive tests themselves (based on raw scores) should be accurate, as they are adjusted for age, race, sex, and education.

The center-to-center differences in the weight given to cognitive tests are not likely to be an artifact of difference in the use of raw versus adjusted cognitive test scores.

Our results indicate that the diagnoses made by consensus versus single clinicians are more likely to be MCI rather than normal (and possibly MCI rather than AD), given the same demographic variables and cognitive test scores. It is possible that diagnoses which are 'hard calls' are more likely to be declared MCI in a consensus process than by a single clinician. It should be noted that we were unable to control for the possible confounding effects of ADC in our analysis of consensus diagnosis; some centers use all or almost all consensus diagnoses, and these centers might also be more likely to diagnose MCI versus normal, for example, than other centers.

There are several other limitations to our findings. It is likely that there are other variables, not included in the UDS, which are used by different centers to determine diagnosis, and which contributed to intercenter variability. One example would be additional cognitive tests (e.g. a test for visuospatial ability). Another unknown factor is the source of subjects in each center, who may come as volunteers from health fairs or relatives of patients, for example. Differences in such groups, not captured by demographic variables in our analysis, could confound our analysis. Finally, it should be acknowledged that our results apply to the ADCs only, and may or may not have applicability to general clinical practice ('external validity').

Nonetheless, our findings suggest the existence of process-dependent and site-specific influences that impact diagnoses. These influences introduce an important source of variability that may adversely affect research studies that depend on accurate and consistent diagnostic classification of research participants. In light of these findings, we believe it would be valuable for NACC to create a test data set of hypothetical subjects, which could be circulated to all ADCs to explore the sources of variation in diagnostic criteria between centers. At the same time, NACC could conduct a survey of how cognitive tests are currently adjusted for age, sex, race, and education across centers. Furthermore, the test data might specify a method of normalizing cognitive tests for age, sex, race, and education.

Regarding the relative importance of predictor variables, it is perhaps not surprising that the most important would be clinician-reported decline and CDR sum of boxes. The relative importance of cognitive tests, with the MMSE in the lead, might also to be expected. As noted, the role played by education, race, and age, might be expected if centers are not adjusting cognitive tests for these variables, whereby younger age, white race, and more education all significantly increase the risk of receiving a worse diagnosis, given the same levels of all other variables in the model.

Finally, we found that after eliminating clinician judgment predictor variables (CDR sum of boxes, clinician-reported decline, clinician-reported depression, and consensus/nonconsensus diagnosis), a model primarily based on demographic and cognitive tests was able to predict final diagnosis with 78% accuracy, nearly as good as the model including clinician judgment variables. Of particular importance in the model without clinician judgment variables was the FAQ, which played a little role in the presence of clinician judgment variables. The FAQ captures key information about functional abilities that drive clinical judgment.

Appendix 1: Description of Tests in Table 1

Cognitive Tests

Boston Naming Test. Assesses the ability to name pictures of objects (30 items used in UDS). In UDS scored as total objects spontaneously named in 20 s plus number named with a semantic cue (e.g. 'It's found in Egypt') for items that are misperceived or that the person does not give any response at all (e.g. 'I don't know what that is'). Tests naming (language).

Digit Symbol Test. Consists of digit-symbol pairs (e.g. 1/-,2/⊥... 7/Λ,8/X,9/=) followed by a list of digits. Under each digit, the subject should write down the corresponding symbol as fast as possible. The number of correct symbols within 90 s (UDS version) is measured. Tests visuomotor processing speed.

MMSE Test. Includes simple questions and problems in a number of areas: the time and place of the test, repeating 3 words, sustained attention such as counting by serial sevens, language use and comprehension, and basic motor skills. For example, one question asks to copy a drawing of two pentagons. Tests overall cognitive status.

Trail Making Tests A and B. Neuropsychological tests of visual attention and task switching. The task requires a subject to 'connect the dots' of 25 consecutive targets on a sheet of paper or computer screen. Two versions are available: A, in which the targets are all numbers (1, 2, 3, etc.), and B, in which the subject alternates between numbers and letters (1, A, 2, B, etc.). The goal of the subject is to finish the test as quickly as possible, and the time taken

to complete the test is used as the primary performance metric. Test visuomotor processing speed (A) and executive function (B).

Digit Span Forward and Backward Tests. Digit Span Tests assess verbal attention and working memory. The person hears a string of numbers and either repeats them in the same order (Digit Span Forward) or reverses the order (Digit Span Backwards).

Logical Memory Test. Assesses verbal memory by asking the subject to recall a story immediately following an oral presentation (part I) and again after at least a 20-min delay (part II). Scored as total items recalled. Tests memory.

Category Fluency Tests. Ask the subject to name as many items in a particular category (animal and vegetables are used in UDS) in a specified time frame (1 min per category). Tests verbal fluency (language).

Clinical Assessment Tests

Clinical Dementia Rating. A semi-structured interview with both the patient and a reliable informant to rate performance in memory, orientation, judgment and problem solving, community affairs, home and hobbies, and personal care. Each area is rated according to 1 of 5 levels of impairment: 0 = none, 0.5 = questionable, 1 = mild, 2 = moderate, 3 = severe. Is either scored with a global summary score (0–3), or with the CDR sum of boxes (0–18), higher scores being worse.

Geriatric Depression Scale. A 15-item scale to measure depression, higher scores indicate more depression.

Hachinski Ischemia Score. A clinical evaluation which differentiates vascular dementia from degenerative forms of the disorder. Scored via 13 items, with scores ranging from 0 to 18; patients with a score of 7 or higher are more likely to have a vascular dementia.

Neuropsychiatric Inventory Questionnaire Summary Score. Based on a structured interview with an informant who is familiar with the subject. It includes questions on a number of neuropsychiatric domains: delusions, hallucinations, dysphoria, anxiety, agitation/aggression, euphoria, disinhibition, irritability/lability, apathy, aberrant motor activity, and night-time behavior disturbances. Absence/presence, as well as frequency and severity are scored.

Functional Assessment Questionnaire. A series of 10 questions about functional ability given to an informant familiar with the subject. Questions cover check balancing, shopping, and food preparation, among others. Each function rated from 0 (no problem) to 3 (dependent on others).

Acknowledgments

Felicia Goldstein provided valuable advice regarding neuropsychological tests. This work was supported by an NIH-NIA Center Grant for the Emory Alzheimer's Disease Research Center (P50AG025688).

References

1 Morris JC, Weintraub S, Chui HC, Cummings J, Decarli C, Ferris S, Foster NL, Galasko D, Graff-Radford N, Peskind ER, Beekly D, Ramos EM, Kukull WA: The Uniform Data Set (UDS): clinical and cognitive variables and descriptive data from Alzheimer Disease Centers. Alzheimer Dis Assoc Disord 2006;20:210–216.

2 Beekly DL, Ramos EM, Lee WW, Deitrich WD, Jacka ME, Wu J, Hubbard JL, Koepsell TD, Morris JC, Kukull WA, NIA Alzheimer's Disease Centers: The National Alzheimer's Coordinating Center (NACC) database: the Uniform Data Set. Alzheimer Dis Assoc Disord 2007;21:249–258.

3 Petersen RC: Mild cognitive impairment as a diagnostic entity. J Intern Med 2004;256: 183–194.

4 Stephan BC, Matthews FE, McKeith IG, Bond J, Brayne C, Medical Research Council Cognitive Function and Aging Study: Early cognitive change in the general population: how do different definitions work? J Am Geriatr Soc 2007;55:1534–1540.

5 Weintraub S, Salmon D, Mercaldo N, Ferris S, Graff-Radford NR, Chui H, Cummings J, DeCarli C, Foster NL, Galasko D, Peskind E, Dietrich W, Beekly DL, Kukull WA, Morris JC: The Alzheimer's Disease Centers' Uniform Data Set (UDS): the neuropsychologic test battery. Alzheimer Dis Assoc Disord 2009;23:91–101.

6 Geisser S: The predictive sample reuse method with applications. J Am Stat Assoc 1975; 70:320–328.

7 Stone M: Cross-validatory choice and assessment of statistical predictions. J R Stat Soc Series B Stat Methodol 1974;36:111–147.

8 Steyerberg EW, Harrell FE Jr, Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD: Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol 2001;54: 774–781.

9 Little J, Rubin D: Statistical Analysis with Missing Data, ed 2. Hoboken, Wiley, 2002.

10 Van Buuren S, Oudshoorn C: Mice: multivariate imputation by chained equations. R package version 1.16. 2007.

11 R Core Development Team: R: A Language and Environment for Statistical Computing. Vienna, R Foundation for Statistical Computing, 2008.