Ascertainment Bias in the Clinical Diagnosis of Alzheimer Disease

Martha Storandt, PhD; John C. Morris, MD

Background: The clinical diagnosis of Alzheimer disease (AD) is often based, at least in part, on poor cognitive test performance compared with normative values.

Objective: To examine the presence and extent of an ascertainment bias (omission of affected individuals) produced by such criteria when applied as early as possible in the course of the disease.

Design: Longitudinal study (1979-2008).

Setting: Washington University Alzheimer Disease Research Center, St Louis, Missouri.

Participants: Of 78 individuals aged 65 to 101 years enrolled as healthy controls, 55 later developed autopsy-confirmed AD; 23 remained cognitively healthy and did not have neuropathologic AD.

Main Outcome Measures: Criteria for the diagnosis of AD based on various cutoff points (1.5, 1.0, and 0.5 SDs below the mean for robust test norms) for 2 standard psychometric measures from each of 3 cognitive domains (episodic memory, visuospatial ability, and working memory) were applied to data from the first assessment associated with an independent clinical diagnosis of cognitive impairment for those who developed symptomatic AD and from the last assessment for those who did not.

Results: Areas under the curve from receiver operating characteristic analyses ranged from 0.71 to 0.49; sensitivities and specificities were unsatisfactory even after adjusting for age and education, using combinations of tests, or examining longitudinal decline before clinical diagnosis.

Conclusion: Reliance on divergence from group normative values to determine initial cognitive decline caused by AD results in failure to include people in the initial symptomatic stage of the illness.

Arch Neurol. 2010;67(11):1364-1369

To implement effective treatment, when available, it is important to identify individuals as early as possible in the neuropathologic course of Alzheimer disease (AD). Clinical diagnosis often relies on comparison of performance on cognitive tests with group norms. In their seminal article, Blessed et al1 reported that some cases had substantial neuropathologic abnormalities at autopsy but performed well on measures of cognition, an observation confirmed by other researchers.2-5 Blessed et al1(p807) suggested that “a certain amount of the change estimated by plaque counts may be accommodated within the reserve capacity of the cerebrum without causing manifest intellectual impairment.” Thus, individuals with greater biological cerebral reserve could experience extensive neuropathologic abnormalities but not reach the threshold for expression of symptoms necessary for clinical diagnosis because of that reserve capacity.6 This hypothesis, later expanded to include cognitive reserve (eg, experience and strategies),7 has been popular in research on AD.

Another obstacle to early diagnosis is ascertainment bias,8 which focuses on methodological issues related to determining the clinical diagnosis of dementia. For example, some popular methods for identifying dementia were designed to detect frank dementia and are too easy (ceiling effect) for high-functioning individuals in the beginning stages of a dementing illness. In one meta-analysis,9 6 of the 13 brain reserve studies used such mental status instruments as the only outcome measure. Even if instruments are sufficiently difficult, cutoff points used to diagnose dementia are usually based on norms from samples that contain individuals with early cognitive impairment whose deficits have not yet been identified as dementia.10 This produces a cutoff point that is too lenient, and all individuals will be further along in the disease pro-
cess when they are identified. For example, in the Nun Study, only 59% of those who met the neuropathologic criteria for AD also met the clinical criteria for dementia, although almost all had some detectable cognitive deficit.

The objective of this study was to assess whether there is ascertainment bias produced by test criteria similar to those often used for the clinical diagnosis of AD when they are applied as early in the course of the disease as possible. The sample included individuals from a longitudinal study who were cognitively healthy at enrollment but later developed autopsy-confirmed symptomatic AD. The sensitivity of the test criteria was examined at the time of first clinical diagnosis of a deficit in cognitive function that reflected intraindividual change as elicited by interviews with informants and the individuals rather than by the person’s relative standing based on robust group norms or deviation from predicted scores based on age and education. Although the primary focus was on the percentage of persons with symptomatic AD who could be identified using test scores, individuals from the same sample who did not develop clinical symptoms of cognitive deficits and who did not have a diagnosis of any dementing disease at autopsy were included as a control group. Longitudinal course was also examined using growth and survival analyses.

METHODS

PARTICIPANTS

Archival data from 78 participants (36 men; 2 black individuals and the remainder white) were available from a sample of 521 volunteers aged 65 to 101 years at entry and enrolled between 1979 and 2008 in a longitudinal study of healthy aging and dementia at the Washington University Alzheimer Disease Research Center. At entry, all the participants were clinically evaluated to be cognitively healthy (Clinical Dementia Rating (CDR) 0), with no potentially confounding neurologic or psychiatric disorders. Of 163 people who progressed to a CDR greater than 0 at their last assessment, 96 died, 63 had autopsies, and 55 (included herein as the AD group) had neuropathologically confirmed AD. The 8 individuals omitted had other neuropathologic diagnoses (5 had a clinical diagnosis of symptomatic AD and 3 were diagnosed as having a non-AD dementia), yielding a clinical diagnostic accuracy rate for AD of 92% (55 of 60 participants). The control group included the 23 people who were still CDR 0 at their last assessment before death and for whom no dementing disease was found at autopsy.

Sample demographics and mean scores on a brief cognitive screening measure are given in Table 1. The percentage of people with an apolipoprotein E4 (APOE4) allele was comparable in the AD (26%) and control (19%) groups (P = .50). The Washington University Human Studies Committee approved all the procedures. Data from these participants have appeared in other publications from the Alzheimer Disease Research Center.

CLINICAL EVALUATION

The CDR staging and diagnostic evaluation is based on annual semistructured interviews with the participant and a knowledgeable collateral source (usually the spouse or an adult child), a health history, medication and depression inventories, an aphasia battery, and a neurologic examination of the participant. Generally, participants were seen by different clinicians from year to year, although the clinician was the same at the next visit for 17% of the follow-up visits (91 of 546). Clinicians did not have access to previous clinical evaluations or to previous and current psychometric test results. The research-trained clinician determined whether any cognitive problems represented decline from former level of function for that individual and interfered to some degree with the person’s ability to perform accustomed activities. If so, a CDR greater than 0 was assigned. This procedure is highly successful (93%) in identifying autopsy-confirmed AD, even at the earliest symptomatic stage that elsewhere is characterized as mild cognitive impairment.

NEUROPATHOLOGIC DIAGNOSIS

The neuropathologic diagnosis of AD was made according to standard assessment procedures. To maintain consistency of diagnosis across time, given that participants were enrolled beginning in 1979, the criteria proposed by Kachaturian and modified by McKee et al were used for diagnosis.

PSYCHOMETRIC ASSESSMENT

A psychometric battery assessing a broad spectrum of abilities was administered to all the participants, usually a week or 2 after the annual clinical assessment. The standard tests examined herein are ones often used in clinical diagnosis based on normative values and included 2 measures from each of 3 cognitive domains, episodic memory, speeded visuospatial ability, and working memory. The episodic memory tests were immediate recall of Logical Memory (verbatim scored according to the criteria by Russell and Associate Learning from the Wechsler Memory Scale (WMS). The visuospatial tests were Block Design and Digit Symbol from the Wechsler Adult Intelligence Scale (WAIS). The tests of working memory were WMS Digit Span Backward and Letter Fluency for /s/ and /p/.

Test scores were converted to 2 scores using as the reference (normative) group a sample of 310 people (mean [SD] age, 74.3 [8.6] years; mean [SD] educational level, 14.8 [3.2] years), including the 23 in the normal brain group reported herein, who similarly were enrolled as CDR 0 during the same period and had at least 1 annual follow-up visit but never progressed to CDR greater than 0. The mean (SD) scores for the measures were as follows: Logical Memory, 8.87 (2.81); Associate Learning, 13.42 (3.53); Block Design, 30.05 (8.63); Digit Symbol, 45.67 (11.53); Digit Span Backward, 4.75 (1.28); and Letter Fluency, 29.41 (9.73).

Beginning September 1, 2005, the funding agency required changes in 3 tests: WMS Logical Memory was replaced by
cutoff points based on group norms, residual scores, or last assessment (control group). Classification accuracy of the APOE4 association for the presence of the APOE4 allele was examined for each cognitive measure based on group norms. The areas under the curve from receiver operating characteristic analyses using group norms and individual residuals were poor (Table 2), as was sensitivity (percentage of correct classifications in the AD group) (Table 3). The commonly used cutoff points on measures of episodic memory at 1.5 or 1.0 SD below the mean produced clearly unacceptable detection rates (23%-44%), although specificity (accurate classification of the control group) was good to excellent (83%-96%). Sensitivity decreased using the individual residuals; specificity increased slightly (Table 3).

The last assessment for the control group. Thus, any significant change in slope applies only to the AD group. Classification accuracy was calculated for a decline in performance of at least 0.5 SD based on the robust group norms between the first CDR greater than 0 and the assessment just before that for the AD group and between the last 2 assessments for the control group.

The areas under the curve from the receiver operating characteristic analyses using group norms and individual residuals were poor (Table 2), as was sensitivity (percentage of correct classifications in the AD group) (Table 3). The commonly used cutoff points on measures of episodic memory at 1.5 or 1.0 SD below the mean produced clearly unacceptable detection rates (23%-44%), although specificity (accurate classification of the control group) was good to excellent (83%-96%). Sensitivity decreased using the individual residuals; specificity increased slightly (Table 3).

Better sensitivity was obtained when a deficit was determined to occur on any 1 or more of the 6 tests, but specificity deteriorated (Table 3). For example, 81% of the AD group had at least 1 value 1.0 SD below the mean using group norms, but only 33% of the control group had at least 1 value 1.0 SD below the mean using group norms.
was correctly classified. Although specificity improved (65%) if one required a 1.0 SD deficit on at least 2 measures, sensitivity declined to 62%. In approximately one-fourth of the AD group, the cognitive area(s) affected did not include episodic memory for 22% (group norms) to 26% (residuals) of the AD group.

Median time to a CDR greater than 0 for the AD group was 5.0 years (95% confidence interval, 3.6-6.2 years). When age at entry, APOE4 status, education, and z scores at entry on the 6 cognitive measures were entered as the sole covariates in Cox regression analyses, 4 were significant: age, APOE4 status, Logical Memory, and Digit Symbol (Table 4). Only Digit Symbol made a unique contribution when these covariates were entered simultaneously (Table 4). It seems that the variance in time to a CDR greater than 0 explained by age, APOE4 status, and initial score on Logical Memory represented shared variance. Age at entry was correlated −0.37 with APOE4 status, −0.34 with Logical Memory, and −0.72 with Digit Symbol.

The longitudinal slopes of the 6 cognitive measures in the control group were similar to those in the AD group before clinical diagnosis; differences between the slopes approached significance only for Logical Memory (P = .12) (Table 5). The significant inflection points indicate a change in the slopes of the AD group at the time of diagnosis on all measures.

Declines of at least 0.5 SD between their last 2 assessments (85%) of those in the control group also had such a decline if the criterion was a decline of at least 0.5 SD on any of the 6 measures, but a similar percentage (85%) of those in the control group also had such a decline in the slopes of the AD group at the time of diagnosis on all measures. Only Digit Symbol made a unique contribution when these covariates were entered simultaneously (Table 4). It seems that the variance in time to a CDR greater than 0 and the previous assessment in the AD group produced poor sensitivity (50%) for the individual measures (Table 6). It improved (89%) if the criterion was a decline of at least 0.5 SD on any of the 6 measures, but a similar percentage (85%) of those in the control group also had such a decline in the slopes of the AD group at the time of diagnosis on all measures.

<table>
<thead>
<tr>
<th>Table 4. Regression Coefficients for Covariates Entered Singly and Simultaneously in Cox Regression Analyses in the AD Group</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Covariate</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Age</td>
</tr>
<tr>
<td>APOE4</td>
</tr>
<tr>
<td>Education</td>
</tr>
<tr>
<td>Logical Memory</td>
</tr>
<tr>
<td>Associate Learning</td>
</tr>
<tr>
<td>Block Design</td>
</tr>
<tr>
<td>Digit Symbol</td>
</tr>
<tr>
<td>Digit Span Backward</td>
</tr>
<tr>
<td>Letter Fluency</td>
</tr>
</tbody>
</table>

Abbreviations: AD, Alzheimer disease; APOE4, apolipoprotein E4.

The inflection point is P < .001 for all.

<table>
<thead>
<tr>
<th>Table 5. Annual Rate of Change in the Control Group and in the AD Group Before and After Clinical Diagnosis</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Measure</strong></td>
</tr>
<tr>
<td>Logical Memory</td>
</tr>
<tr>
<td>Associate Learning</td>
</tr>
<tr>
<td>Block Design</td>
</tr>
<tr>
<td>Digit Symbol</td>
</tr>
<tr>
<td>Digit Span Backward</td>
</tr>
<tr>
<td>Letter Fluency</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 6. Sensitivity and Specificity Based on a 0.5-SD Decline Between 2 Assessments in the AD and Control Groups and 1-Year Test-Retest Reliabilities in the Robust Norms Reference Group</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Measure</strong></td>
</tr>
<tr>
<td>Logical Memory</td>
</tr>
<tr>
<td>Associate Learning</td>
</tr>
<tr>
<td>Block Design</td>
</tr>
<tr>
<td>Digit Symbol</td>
</tr>
<tr>
<td>Digit Span Backward</td>
</tr>
<tr>
<td>Letter Fluency</td>
</tr>
<tr>
<td>Any measure</td>
</tr>
</tbody>
</table>

Abbreviations: AD, Alzheimer disease; CI, confidence interval; ellipsis, not applicable.

A sensitivity of the percentage of correct classifications in the AD group; specificity, the percentage of correct classifications in the control group.

Divergence from group normative values to determine initial cognitive decline in AD is not effective even with robust norms on measures with sufficient range so as to avoid ceiling effects. At the time individuals were identified clinically as no longer cognitively healthy, commonly used cutoff points such as −1.5 and −1.0 SD on measures of episodic memory detected only 23% to 44% of those later confirmed to have neuropathologic AD. Adjusting for age and education was somewhat useful in preventing cognitively intact people from being labeled as symptomatic AD, particularly on one of the speeded visuospatial measures. Sensitivity, however, decreased, making it more difficult to identify those with symptomatic AD by lowering the cutoff point for older people, who are also more likely to have AD.

The “best” balance between sensitivity (0.72) and specificity (0.61) was obtained when a −1.0 SD criterion based on residuals was met for any of the 6 measures. Even using information from any 3 of the cognitive domains affected by AD, using robust norms, and adjusting for age and education, a diagnosis based on objective test performance at that time was wrong almost a third of the time.

In addition to failing to detect symptomatic AD as early as possible (including individuals characterized elsewhere as having mild cognitive impairment) for treatment purposes, research samples determined by such normative criteria do not include people in the initial...
symptomatic stage of the disease. Similar to the problem of contamination of putatively normal comparison samples pointed out previously, samples are also biased if they omit people in the beginning phases of symptomatic AD. Depending on the purpose of the research, results may be misleading. For example, it would be difficult to identify the sequence of pathologic processes that occur at disease onset if individuals in the very early stage are absent.

The survival analyses suggest that a speeded visuospatial test (Digit Symbol) is a stronger predictor of time to clinical detection of AD than is age, APOE status, or episodic memory. Block Design, another visuospatial test, did not demonstrate the same effect. Digit Symbol reflects perceptual speed, whereas Block Design, although timed, is primarily a measure of spatial ability. It is possible that perceptual speed is affected early by the pathologic process, but the comparable rates of decline on Digit Symbol in the symptomatic AD and control groups before clinical diagnosis of AD would argue against that interpretation. Alternatively, because processing speed is strongly related to fluid intelligence, perhaps it would be a better indicator of brain or cognitive reserve than is educational level, which was unrelated to either the development of symptomatic AD or the time to clinical diagnosis in this sample.

We have argued that it is necessary to consider the person’s longitudinal course rather than emphasizing current standing in a normative group. Although the present results illustrate the limitations of emphasis on the person’s current standing, it is not clear that applying a longitudinal approach to standard psychometric tests provides a satisfactory alternative. The linear rate of decline in symptomatic AD (Table 5) is substantial during the course of the disease, but it is actually an accelerating rate of decline. For standard cognitive measures, such as the ones used herein, the annual decline is less in the early symptomatic phases of the disease than it is later, making it difficult to differentiate cognitive decline from measurement error. Logical Memory, for example, dropped 0.5 SD in approximately half of both groups (Table 6). Test-retest stability may be inadequate for a variety of standard measures. To examine that possibility, we calculated 1-year test-retest reliabilities (Table 6) in the reference sample of cognitively healthy individuals using values obtained at their second and third assessments so as to avoid the practice effect often seen after the first assessment. Half the measures had poor (<0.70) test-retest stability.

Although this study has several strengths (eg, autopsy confirmation of AD diagnosis, initial cognitively healthy status of all participants, robust norms, and statistical correction for age and educational level), it is a study of research volunteers, which limits generalization. A larger sample, especially of the autopsy-verified control group, would increase statistical power, which might allow detection of memory decline before clinical diagnosis. The cognitive measures described herein are widely used, but they represent a limited set of available instruments; furthermore, they do not include measures of delayed list-learning recall or of executive function, attention, and working memory developed since this longitudinal research began. If cognitive measures are to be used in the clinical diagnosis of AD when it first occurs, more sensitive, reliable measures of memory and other cognitive domains affected by the disease are needed to avoid the ascertainment bias that currently affects many research samples.

Clinicians must also carefully assess changes in function as determined by history or collateral report, especially in initially high-functioning individuals. Comparatively little attention has been given to the development of measures of subtle changes in function for either clinical practice or research. The AD8 is one such attempt; it is a brief (8-question) informant interview that is sensitive to early features of cognitive change and is highly correlated with cognitive and functional components of the CDR and with neuropsychological testing.

Accepted for Publication: February 10, 2010.
Correspondence: John C. Morris, MD, Alzheimer Disease Research Center, Washington University, 4488 Forest Park Ave, Ste 130, St Louis, MO 63108 (morrisj@abraxas.wustl.edu).

Author Contributions: Study concept and design: Storandt. Acquisition of data: Storandt and Morris. Analysis and interpretation of data: Storandt and Morris. Drafting of the manuscript: Storandt. Critical revision of the manuscript for important intellectual content: Morris. Statistical analysis: Storandt. Obtained funding: Storandt and Morris. Administrative, technical, and material support: Morris.

Financial Disclosure: None reported.

Funding/Support: This study was supported by grants P01 AG03991 and P50 AG05681 from the National Institute on Aging.

Additional Contributions: We thank Daniel W. McKeel Jr, MD, and the Washington University Alzheimer Disease Research Center Neuropsychology Core (Nigel Cairns, PhD, core leader) for the neuropathologic diagnoses and the Genetics Core (Alison Goate, DPhil, core leader) for APOE status.

REFERENCES

10. Sliwinski M, Lipton RB, Buschke H, Steward W. The effects of preclinical demen-

Announcement

New Initiatives: Clinical Trials and Videos. We have embarked on 2 new initiatives: Clinical Trials and video presentations. We welcome manuscripts that describe double-blind, randomized, placebo-controlled clinical trials as our primary area of interest. We plan on expediting the review process and time to publication and to include them online ahead of print as these studies are time sensitive and of direct benefit to our patients. We hope you will take advantage of this new initiative. Please refer to the Instructions for Authors when submitting a Clinical Trials paper, including the requirement to register the trial with an accepted clinical trials site.

We plan to utilize videos as part of published papers that highlight and provide convincing information about the observational and visual features of a patient’s neurologic findings. Please refer to Instructions for Authors for instructions on submitting video presentations.