Evaluating the Quality of Longitudinal Statistical Applications in Original Publications on Alzheimer's Disease

Chengjie Xiong^{a, b} Yuxiao Tang^c Gerald van Belle^d J. Philip Miller^a
Lenore J. Launer^e John C. Morris^{b, f}

^aDivision of Biostatistics and ^bAlzheimer's Disease Research Center, Washington University, St. Louis, Mo.,

^cProgram for Appropriate Technology in Health, ^dDepartment of Biostatistics, University of Washington, Seattle, Wash.,

^eLaboratory of Epidemiology, Demography, and Biometry, National Institute on Aging, Maryland, Md., and

^fDepartments of Neurology, and Pathology and Immunology, Washington University, St. Louis, Mo., USA

Key Words

Alzheimer's disease, longitudinal studies ·
Neurodegenerative disease · Cognition ·
Reliability coefficient

Abstract

Background/Aims: To evaluate the quality of longitudinal statistical applications in published studies on Alzheimer's disease (AD). **Methods:** A 21-item instrument, the Quality of Longitudinal AD Studies (QLADS), was developed by the research team (4 biostatisticians, 1 neuroepidemiologist, and 1 neurologist). All items were extensively discussed within the team for content validity. After pilot testing on 5 publications, the instrument was revised and tested for reliability with a sample of 40 published longitudinal AD studies randomly sampled from MEDLINE. **Results:** Item-specific test-retest reliability coefficients for QLADS ranged from 0.53 to 1.00 with the associated standard error (SE) ranging from 0.02 to 0.13. The test-retest reliability for the overall score over the 21 items was high (intraclass correlation coefficient (ICC) = 0.94, 95% CI 0.90, 0.97). Item-specific inter-rater reliability coefficients for QLADS ranged from 0.46 to 1.00 with

the associated SE ranging from 0.07 to 0.18. The inter-rater reliability for the overall score was also high (ICC = 0.87, 95% CI 0.77, 0.93). **Conclusions:** This study indicates that the quality of longitudinal statistical applications in AD publications can be reliably assessed.

Copyright © 2008 S. Karger AG, Basel

Introduction

Alzheimer's disease (AD) is a neurodegenerative disease that results in the irreversible loss of neurons in the cerebral cortex and hippocampus [1]. AD is characterized clinically by memory loss in early stages and increasing severe debilitating symptoms as the disease progresses over what could be as long as a 20-year period. The defining characteristic of AD is its progressive cognitive decline and deteriorating functional abilities, which can only be captured by longitudinal follow-ups on elderly individuals. The natural history of normal aging and AD requires longitudinal examinations of elderly individuals through repeated assessments on cognition, clinical symptoms, cerebral spinal fluid biomarkers, and neuro-

imaging markers. The progressive nature of AD mandates the longitudinal predictions to critical disease milestones, such as age at disease onset, nursing home placement, and death, whereas a cross-sectional prediction may at times even offer a biased estimate to the risk of AD and therefore result in inaccurate predictions. This is highlighted by a recent study that examined the cognition on high-functioning people for preclinical signs of AD [2] and the prediction to future development of mild cognitive impairment (MCI) [3] or AD.

Longitudinal studies are those that investigate changes over time, possibly in relation to an intervention. The primary characteristic of a longitudinal study is that study subjects are measured repeatedly through time. Outcome variables in the longitudinal studies may be continuous measurements, counts, dichotomous or categorical indicators, and in many cases, outcomes may be multivariate as well [4–8]. Additional design and analytic features of a longitudinal study include, but are not limited to, intra-subject correlations among repeated measures from the same subjects, baseline differences and individual level characteristics as potential confounding covariates to the rate of change over time, missing data and censoring observations. On the other hand, there are rather unique design and analytic implications associated with biomedical and clinical studies of AD. These specific features include, but are not limited to, the use of large psychometric batteries, possible floor and ceiling effects in psychometric batteries, typically non-linear progression patterns over time on cognitive and functional measures, and demographics (e.g., age, education, and gender) as important risk factors of AD.

We recently reported several instruments/checklists that were designed to systematically assess the types of statistical designs and analytic methods, and the quality of statistical reporting in published AD studies [9]. These checklists were intended to offer an understanding of the statistical characteristics common to all or most AD studies. Our goal in the current report is to develop a reliable and valid instrument to evaluate the quality of longitudinal statistical applications in publications on AD and to survey the types of statistical methods used.

Materials and Methods

Development of Pilot Instrument

We were primarily interested in two types of longitudinal studies which covered essentially all published longitudinal AD studies. The first one used repeated measures of disease markers to track the longitudinal courses of AD progression, and the sec-

ond one employed clinically important disease milestones (e.g., AD onset, nursing home placement, and death) to track the progression of AD by measuring the time and the speed from the baseline to the development of these milestones. Both types of longitudinal studies required specific types of statistical models which were valid only under certain distributional and regularity conditions, and the implementations of these models required complicated statistical computing techniques. Both types of longitudinal studies in AD were mostly observational in nature [9], which usually implied variable baseline conditions and unequal follow-up times among individuals. Dropouts and loss to follow-up usually occurred in longitudinal studies of AD, resulting in either missing or censored data which were likely informative and not ignorable [10–13]. Baseline patient characteristics such as the well-known risk factors of AD (i.e., age, apolipoprotein E4, gender, education) and other time-varying covariates could affect the longitudinal rate of change in both types of longitudinal studies.

We conceptualized two distinct statistical components in a published original longitudinal AD study. The first component contained the statistical features that were shared by almost all AD studies. These statistical aspects were the main targets of our previous report and can be assessed in AD publications by the Assessment of Statistical Reporting [9]. The second statistical component contained statistical features specific to longitudinal studies which, when coupled with the specific analytic features associated with AD, had rather unique and complicated analytic implications in longitudinal AD studies and were the primary focus of this study. We developed a pilot instrument, Quality of Longitudinal AD Studies (QLADS), to assess how these unique design and analytic features were dealt with and reported in published longitudinal AD studies. For example, the first item in the instrument assessed whether longitudinal statistical models were used to analyze longitudinal data, which is the most fundamental analytic difference between longitudinal studies and cross-sectional studies. Similar items included whether appropriate modules from well-validated statistical computing packages (e.g., SAS, SPSS, and STATA) were used to implement these analyses. There were also items that assessed whether the goodness-of-fit for the longitudinal models was examined, which is crucial as it has been well established in the literature that the longitudinal progressive course of AD is not linear and depends on the stage of disease severity [14–16]. Differential baseline features [17–20] and length of follow-up could potentially affect the outcome of longitudinal studies, and several items in this instrument assessed whether these features were presented and compared among different study groups. Missing data were common in essentially all longitudinal AD studies. Several items in the instrument examined the differing amount of missing data among study groups, the difference between those who completed the study and those who dropped out on basic demographic characteristics, and a report on missing and censoring data mechanisms as well as the possible effects on statistical inferences. Many AD studies employed multiple longitudinally assessed disease markers (e.g., multiple cognitive tests) or multiple clinically significant endpoints (e.g., these based on conversion to MCI [3] and to different stages of dementia as measured by the Clinical Dementia Rating [21]). One item assessed whether such inter-correlations among multiple outcome variables were taken into account in statistical analyses.

In addition to developing QLADS, we developed two independent items to specifically assess the types of outcome variables

and the types of statistical models used in these analyses. The first item read, 'If longitudinal statistical analyses were reported in the Abstract, which types of variables were studied?' The answer to this item was one of the following four mutually exclusive options: 1 = repeated measures type only; 2 = time-to-event type only; 3 = both 1 and 2, and 4 = neither 1 nor 2. The second item read, 'If longitudinal statistical analyses were reported in the Abstract, which statistical models were used?' The answer to this item was one of the following options: 0 = paired sample t tests or cross-sectional analyses based on appropriate summary measures calculated over longitudinal observations from each subject when the study design is complete and balanced (i.e., all subjects measured at the same two time points); 1 = only cross-sectional analyses were done either at each time point or based on inappropriate summary measures calculated over longitudinal observations from each subject (i.e., unbalanced and incomplete designs); 2 = repeated measures ANOVA or ANOCOVA; 3 = general linear models for longitudinal data; 4 = generalized linear models for longitudinal data (including transition models of Markov chain type); 5 = nonlinear mixed effects models; 6 = distribution-free tests for comparing survival distributions; 7 = proportional hazards models; 8 = accelerated failure time models; 9 = frailty models; 10 = epidemiologic methods/models (incidence/prevalence rates, mortality, morbidity, etc.), and 11 = others. Because some ratings (e.g., 2, 3, and 4) were nested (i.e., previous ones can be regarded as special cases of later ones), if the rating to an article was multiple, the lower level choice (i.e., the smaller numerical code) was chosen.

Content validity of all items in QLADS was extensively discussed among the study team members. Because items in QLADS assessed unique design and analytic features resulting from the combination of longitudinal study designs and AD, their content validity was based on the fact that these specific statistical features were so essential in longitudinal AD studies that any statistical designs and analyses ignoring them could result in biased longitudinal statistical inferences and led to unjustified scientific conclusions. For example, it is well known that if repeated measures were analyzed in a cross-sectional way by ignoring the within-subject correlations among repeated measures, the resulting statistical inferences could be biased [4, 22].

Scoring

In QLADS, the rating for each item depended on how well the specific criterion was reported in the publication and was given in a 5-point Likert scale as: excellent (coded as 4), above average (3), average (2), below average (1), and extremely poor (0). We chose a Likert scale to represent the fact that a simple dichotomous scale, as used in many existing instruments for the methodological assessment of controlled clinical trials, might not capture the true amount of variation in published longitudinal AD studies. As an example, for the item assessing whether the goodness-of-fit of the longitudinal statistical models was examined and clearly reported, there could be several ways to do so for each longitudinal statistical model used in the analyses, such as formal statistical tests [4, 23] and model fitting diagnostics as well as various residuals plots [4, 23, 24], necessitating the need of a finer rating than a simple binary rating to the item. As another example, AD studies usually involved multiple outcome measures due to the use of large neuropsychometric batteries, a 5-point Likert scale was appropriate to differentiate whether the goodness-of-fit

assessment on multiple longitudinal models was reported for all (i.e., excellent), or for the majority but not for all (i.e., above average), or for only about half (i.e., average), or for fewer than half but not for none (below average), or for none (i.e., extremely poor) of these psychometric tests. An overall quality score of statistical reporting for each article was obtained by the mean score of 21 items multiplied by 25. This yielded a scale between 0 and 100. A score of 100 represented the best possible quality in longitudinal statistical applications, and a score of 0 represented the worst possible quality score.

Sampling of Longitudinal AD Studies from MEDLINE

A research assistant performed a keyword search using either of the words 'Alzheimer' or 'AD' along with the word 'longitudinal' after receiving appropriate training on the use of MEDLINE database offered by the Medical Library of Washington University School of Medicine. The search was restricted to human subjects, the English language, the years from 1984 to 2005, and the following 14 journals that frequently published AD studies: *American Journal of Epidemiology*; *Annals of Neurology*; *Archives of General Psychiatry*; *Archives of Neurology*; *Alzheimer Disease and Associated Disorders*; *Brain*; *Journal of the American Geriatrics Society*; *Journal of the American Medical Association*; *Lancet*; *Neurobiology of Aging*; *Neuroepidemiology*; *Neurology*; *New England Journal of Medicine*, and *Psychology and Aging*. The target population for this study was the peer-reviewed original publications of longitudinal studies on AD from these 14 journals, and therefore the search also excluded comments, editorials, letters and review articles. These search criteria resulted in a total of 380 published longitudinal AD studies from MEDLINE as of February 10, 2006. Two independent random samples of AD publications (one for the pilot tryout and the other for a formal reliability study) were created using two independent simple random sampling schemes with the random number generator of SAS [25]. Authors' identity, institutional affiliations, publication date, and journal identity were all removed from these articles before they were rated.

Revision of the Pilot Instrument

The pilot instrument QLADS was first tested in a pilot tryout of 5 published longitudinal studies on AD. Two reviewers (C.X., Y.T.) carefully read all 5 articles and discussed all the ratings item by item by phone and E-mails. Items were simplified or dropped if consensus decisions were difficult to reach or if there was insufficient information in these publications, unless the item was considered absolutely essential for the evaluation of study quality by at least one member of the study team. Many items were rephrased to avoid ambiguity while keeping the core longitudinal statistical features for the content validity. The definitions of all ratings were further revised to reflect these changes. The revised instrument contained a total of 21 items as presented in table 2. The detailed item descriptions and the associated rating methods of the instrument can be obtained by E-mail to the first author at chengjie@wubios.wustl.edu.

Testing of the Revised Instrument

The revised instrument QLADS was formally tested for reliability using a simple random sample of 40 published longitudinal studies on AD from MEDLINE. Item-specific inter-rater and test-retest reliability as well as the inter-rater and test-retest reliability

for the overall score on the quality of longitudinal statistical applications in AD studies were the major focus in the testing. For each article, the research assistant made 3 copies. To assess inter-rater reliability, one copy of each was given to each reviewer. To assess intra-rater reliability, the other copy was given to one reviewer (C.X.) about 3 weeks after the same article was rated. The two reviewers had the detailed definitions of item ratings and rated each article independently.

Statistical Analyses

Item-specific inter-rater and test-retest reliability in QLADS was assessed by the weighted Kappa coefficient [26] and the corresponding standard error. For the overall score of QLADS, intra-class correlation coefficients (ICCs) [27] along with the asymptotic 95% CIs [28] were computed for both the inter-rater and test-retest reliability assessments. For the item assessing the types of longitudinal outcome variables used in AD studies, item-specific inter-rater and test-retest reliability was assessed by the simple (i.e., not weighted) Kappa coefficient [29] and the corresponding standard error [29]. Because multiple ratings were possible to the item assessing the types of statistical models used in longitudinal analyses, item-specific inter-rater and test-retest reliability for the item was assessed as a binary rating (i.e., yes or no) for the most frequently reported statistical models in the sample by the simple Kappa coefficient and the corresponding standard error [29]. All study data were entered into a secure SAS database and maintained by the research assistant. The statistical analyses were implemented in SAS version 8.1 [25]. A SAS macro was used to compute the ICCs and their 95% CIs [28].

Results

The 40 published longitudinal studies on AD were distributed as follows: *Annals of Neurology* (n = 2); *Archives of Neurology* (n = 6); *Alzheimer Disease and Associated Disorders* (n = 8); *Brain* (n = 2); *Journal of the American Geriatrics Society* (n = 4); *Journal of the American Medical Association* (n = 2); *Neurobiology of Aging* (n = 3); *Neuroepidemiology* (n = 1); *Neurology* (n = 11), and *Psychology and Aging* (n = 1). Table 1 presents the descriptive statistics on the overall score of QLADS from this sample. Table 2 provides detailed questions for all items in QLADS and shows the item-specific inter-rater and test-retest reliability along with the corresponding standard error (SE). Item-specific test-retest reliability coefficients ranged from 0.53 to 1.00 with the associated SE ranging from 0.02 to 0.13. The test-retest reliability for the overall score over the 21 items was high (ICC = 0.94, 95% CI 0.90, 0.97). Item-specific inter-rater reliability coefficients for QLADS ranged from 0.46 to 1.00 with the associated SE ranging from 0.07 to 0.18. The inter-rater reliability for the overall score was also high (ICC = 0.87, 95% CI 0.77, 0.93).

Table 1. Descriptive statistics on the overall score of QLADS from the sample of published longitudinal AD studies (n = 40)

Rater	Mean	Standard deviation	Range
Rater 1 time 1	32.1	14.1	10.7–59.5
Rater 2 time 1	34.3	16.5	4.8–72.6
Rater 2 time 2	33.1	14.9	7.1–71.4

For the item assessing the types of longitudinal outcome variables used in the longitudinal studies, the inter-rater reliability was 0.95 with an associated SE of 0.05, and the test-retest reliability was 1.00. For the item assessing the types of statistical methods and models used in the longitudinal analyses, the most frequently identified statistical models in the sample of 40 were 1, 6, and 7 (see the codes in Development of Pilot Instrument, above). The inter-rater reliability coefficients for identifying model 1, 6, and 7 were 0.90, 0.83, and 0.84, respectively. The associated SEs for these inter-rater reliability coefficients were 0.07, 0.12, 0.11, respectively. The test-retest reliability coefficients for identifying model 1, 6, and 7 were all 1.00.

Discussion

Longitudinal follow-ups of elderly individuals on disease markers and milestone events play a crucial role in understanding AD and normal aging as well as their differences. There is a growing urgency to find effective disease-modifying therapies to treat AD in current AD research. In order to do so, patients must be longitudinally assessed in order to establish the efficacy of the treatments by comparing the patients receiving the treatments and those receiving the placebo. The current AD research also points to the crucial role of predicting future AD or MCI when individuals are still cognitively normal. An accurate prediction of AD can only be fulfilled when cognitively normal individuals are longitudinally followed with well-designed protocols so that even very subtle clinical and cognitive signs can be detected as early as possible.

To our knowledge, however, there has been no systematic review of the longitudinal statistical applications in published longitudinal AD studies, although many instruments have been developed in the literature for assessing the controlled clinical trials which are in

Table 2. Item-specific reliability coefficients in QLADS

Criterion assessed	Inter-rater reliability (SE)	Test-retest reliability (SE)
Are the longitudinal statistical analyses based on longitudinal statistical models?	0.65 (0.09)	0.90 (0.04)
Is there a graphical representation of longitudinal data or descriptive statistics with the indication of variability?	0.80 (0.08)	0.89 (0.05)
Is there an assessment of goodness-of-fit on the longitudinal statistical models used in the publication?	0.57 (0.11)	0.64 (0.10)
Do the analyses make use of all observed data from all subjects (whether repeated measures or time-to-event data)?	0.61 (0.12)	0.76 (0.07)
Does the publication identify the specific modules or functions from the software package used to implement the longitudinal analyses?	1.00	1.00
Are references or details or appendices given specifically for the longitudinal statistical methods used?	0.78 (0.09)	0.89 (0.06)
Are the baseline and the length of follow-up, as well as the frequency of subject assessments, of the longitudinal study clearly described for each treatment group?	0.53 (0.10)	0.58 (0.08)
Are the baseline and the length of follow-up of the longitudinal study compared among treatment groups?	0.51 (0.12)	0.53 (0.09)
Are the amount of missing data (dropout and loss to follow-up) and the time of dropout/loss to follow-up clearly described for each treatment group?	0.47 (0.10)	0.61 (0.07)
Are the amount of missing data (dropout and loss to follow-up) and the time of dropout/loss to follow-up compared among treatment groups?	0.59 (0.11)	0.67 (0.10)
Are the baseline demographics/dementia status clearly described or summarized for each sub-sample of dropouts and loss to follow-up from each treatment group?	0.76 (0.14)	0.91 (0.07)
Are the baseline demographics/dementia status compared among sub-samples of dropouts or loss to follow-up from different treatment groups?	0.66 (0.13)	0.94 (0.04)
Is the overall sub-sample of dropouts and loss to follow-up compared to the overall sub-sample who completed the study on at least baseline demographics and/or dementia status?	0.74 (0.13)	0.84 (0.12)
Is there an assessment on the Missing Data Mechanism (MDM) or the censoring mechanism?	0.73 (0.18)	0.80 (0.13)
Is there an assessment on the effect of missing or censoring data or truncated data on the reported conclusions?	0.65 (0.16)	0.76 (0.11)
Are the baseline and time-varying covariates that could affect the longitudinal course of primary variables clearly described for each treatment group?	0.51 (0.08)	0.68 (0.07)
Are the baseline and time-varying covariates that could affect the longitudinal course of primary variables compared among the treatment groups?	0.46 (0.10)	0.67 (0.08)
Are there statistical adjustments in the longitudinal analyses for the baseline and time-varying covariates that could affect the longitudinal course of the primary variables?	0.75 (0.10)	0.97 (0.02)
If there are multiple response longitudinal variables (either the repeated measures type or time-to-event type or both) measured on the same subjects in the study, do the statistical analyses accommodate the intra-subject correlations among multiple variables?	0.56 (0.13)	0.75 (0.10)
Is there a power analysis for the longitudinal design which is based on longitudinal statistical models?	1.00	1.00
Do the longitudinal statistical analyses support the longitudinal scientific conclusions made in the Abstract?	0.65 (0.07)	0.86 (0.05)
All reliability coefficients reported are weighted Kappa.		

general longitudinal in nature [30–37]. The fact that the majority of published AD studies were observational makes it necessary to conduct a rigorous evaluation of longitudinal statistical applications in published clinical publications of AD. The current study served as the first step toward achieving this goal. We developed a reliable and valid instrument to assess the unique combination of statistical features between longitudinal studies and AD. The inter-rater and test-retest reliability coefficients on the overall quality score of QLADS were high and comparable to those reported on the assessment of methodological quality of drug studies [38]. We also reported the item-specific reliability coefficients for our instrument in this study. Because each item in QLADS was designed to assess a distinctive feature of longitudinal statistical designs and analyses used in AD publications, the reported item-specific reliability coefficients provided useful information in assessing a wide spectrum of longitudinal statistical features in AD literature. Our item analyses revealed acceptable reliability for essentially all items. The minimum item-specific test-retest reliability and inter-rater reliability were 0.53 and 0.46, respectively. The mean item-specific test-retest reliability and inter-rater reliability were 0.79 and 0.67, respectively.

Missing data occur in almost all longitudinal AD studies, and they cause not only technical difficulties in the analysis of such data, but also deeper conceptual issues as one has to ask why the measurements are missing, and more specifically whether their being missing has any bearing on the practical and scientific objectives to be addressed by the data. Our proposed instrument paid special attention to the issue of missing data because the analyses on missing data are an integral part of longitudinal data analyses and the entire statistical inferences could be invalid if the missing data are not appropriately analyzed [10]. A general treatment of statistical analysis with missing data along with a hierarchy of missing data mechanism (MDM) has been proposed [10]. MDM is classified as missing completely at random, missing at random (MAR), or non-ignorable. The MAR implies that the probability of missingness depends only on the observed data, but not on the missing values. The implication of this result is that, as far as the statistical inferences are concerned, likelihood-based longitudinal statistical methods, as the dominant methods used in AD publications, are still valid as long as the distribution of data satisfies the assumptions under which these methods are justified. When MDM is not MAR, on the other hand, longitudinal statistical inferences need to make assump-

tions on MDM which cannot be verified based on observed data.

Our proposed instrument has important implications for the future longitudinal statistical applications of AD research. First, QLADS applies to essentially all longitudinal studies in AD based on the fact that virtually all studies in our sample were one of these two types that QLADS was designed for. Having established the reliability/validity of QLADS, we plan to perform a rigorous and systematic assessment of longitudinal statistical applications in AD publications. Such an assessment will offer an understanding on the quality of longitudinal statistical designs and methods in published clinical AD studies. Second, in 2005, for the first time since the beginning of the Alzheimer's Disease Centers Program established in 1984 in the United States, all AD centers in the US have begun to enroll and follow patients with a common, standardized research protocol and provide data into the Uniform Data Set (UDS) through the National Alzheimer's Coordinating Center. With annual follow-up evaluations planned, a large and growing nationally representative longitudinal database will soon be accrued for AD researchers. The implementation of the UDS provides a unique and perfect opportunity for clinicians and investigators to explore and address a very wide spectrum of scientific questions about AD and aging that will require substantial longitudinal statistical applications. The reliable and valid instrument we developed, coupled with a subsequent large scale statistical review on longitudinal AD publications using the instrument which we plan to conduct, will provide clinicians and investigators with a comprehensive understanding of the type of longitudinal statistical applications required in AD research and appreciate more the longitudinal statistical applications when analyzing the UDS.

For clinicians who have to follow up their patients longitudinally and assess the patients' rate of disease progression over time, our instrument offers another channel for them to understand the fast growing longitudinal AD literature by providing valuable information on the methodological and statistical quality of the literature. This information tends to be more crucial, especially when clinicians are confronted with conflicting reports in the literature and have to make a clinical decision based on the literature. With some basic understandings of longitudinal data analyses or help from biostatisticians, clinicians can easily apply QLADS themselves. Therefore, the clinicians now have a choice during their decision-making process to put more weight on the literature that exhibited better longitudinal statistical

methodologies. For investigators who are engaged in longitudinal AD clinical research, our instrument may aid them to critically review the published research and generate new scientifically sound hypotheses based on these findings from publications with solid longitudinal statistical methodologies.

Some limitations exist for our current report. First, the major findings of this study were based on a random sample of 40 published longitudinal AD studies, which might not be a sufficiently large sample size for some of the reported asymptotic statistical inferences to be completely valid. Second, we used a simple random sampling scheme to select the random sample from the entire population of longitudinal AD publications from 14 journals. The sample missed 4 journals because of the low total counts of longitudinal AD publications from these journals (the expected number of articles from each of the 4 journals in a sample of size 40 was all less than 1). In fact, the small total counts from each of the 4 journals pre-

vented us from using a stratified sampling scheme to address the journal-specific reliability of QLADS. Although we do not think the missing of 4 journals alters the findings presented in this study, a definite answer to the question whether the reported instrument behaves differently among different journals can only be addressed in the future when the number of longitudinal AD publications becomes reasonably large for all journals.

Acknowledgments

This study was supported by grant K25 AG025189 to C.X. from the National Institute on Aging and by the Alan A. and Edith Wolff Charitable Trust. Financial support for this study was also provided in part by National Institute on Aging grants AG 03991 and AG 05681 for J.P.M. and J.C.M. The authors wish to express their sincere thanks to the editors and referees for their constructive and valuable comments and suggestions which considerably improved the manuscript.

References

- McKhann G, Drachman D, Folstein M, Katzman R, Price D, Stadlan EM: Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurology* 1984;34:939-944.
- Rentz DM, Huh TJ, Faust RR, et al: Use of IC-adjusted norms to predict progressive cognitive decline in highly intelligent older individuals. *Neuropsychology* 2004;18:38-49.
- Peterson RC, Doody R, Kurz A, et al: Current concepts in mild cognitive impairment. *Arch Neurol* 2001;58:1985-1992.
- Diggle PJ, Heagerty P, Liang K-Y, et al: *Analysis of Longitudinal Data*, ed 2. New York, Oxford University Press, 2002.
- Fitzmaurice GM, Laird NM, Ware JH: *Applied Longitudinal Analysis*. Hoboken, Wiley, 2004.
- Verbeke G, Molenberghs G: *Linear Mixed Models for Longitudinal Data*. New York, Springer, 2000.
- Singer JD, Willett JB: *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. New York, Oxford University Press, 2003.
- Molenberghs G, Verbeke G: *Models for Discrete Longitudinal Data*. New York, Springer, 2005.
- Xiong C, Tang Y, van Belle G, Miller JP, Launer LJ, Bergmann KR, Morris JC: Assessing statistical applications in publications on Alzheimer's disease. *Neuroepidemiology* 2007;28:235-245.
- Little RJA, Rubin DB: *Statistical Analysis with Missing Data*, ed 2. New York, Wiley, 2002.
- Wu MC, Bailey KR: Estimation and comparison of changes in the presence of informative right censoring: conditional linear model. *Biometrics* 1988;45:939-955.
- Wu MC, Carroll RJ: Estimation and comparison of changes in the presence of right censoring by modeling the censoring process. *Biometrics* 1988;44:175-188.
- Kalbfleisch JD, Prentice RL: *The Statistical Analysis of Failure Time Data*. New York, Wiley, 1980.
- Storandt M, Grant EA, Miller JP, et al: Rates of progression in mild cognitive impairment and early Alzheimer disease. *Neurology* 2002;59:1034-1041.
- Mohs RC, Schmeidler J, Aryan M: Longitudinal studies of cognitive, functional and behavioural change in patients with Alzheimer's disease. *Stat Med* 2000;19:1401-1409.
- Galasko DR, Gould RL, Abramson IS, Salmon DP: Measuring cognitive change in a cohort of patients with Alzheimer's disease. *Stat Med* 2000;19:1421-1432.
- Zonderman AB, Giambra LM, Arenberg D, Resnick SM, Costa PT Jr, Kawas CH: Changes in immediate visual memory predict cognitive impairment. *Arch Clin Neuropsychol* 1995;10:111-123.
- Tierney MC, Szalai JP, Snow WG, et al: Prediction of probable Alzheimer's disease in memory-impaired patients: a prospective longitudinal study. *Neurology* 1996;46:661-665.
- Marquis S, Moore MM, Howieson DB, et al: Independent predictors of cognitive decline in healthy elderly persons. *Arch Neurol* 2002;59:601-606.
- Launer L: Nonsteroidal anti-inflammatory drug use and the risk for Alzheimer's disease: dissecting the epidemiological evidence. *Drugs* 2003;63:731-739.
- Morris JC: The clinical dementia rating (CDR): current version and scoring rules. *Neurology* 1993;43:2412-2414.
- Xiong C, Zhu K, Yu K, Miller JP: Statistical modeling in biomedical research: longitudinal data analysis; in Rao CR, Miller JP, Rao DC (eds): *Handbook of Statistics*. Amsterdam, Elsevier, 2007, pp 429-463.
- Therneau TM, Grambsch PM: *Modeling Survival Data: Extending the Cox Model*. New York, Springer, 2000.
- Schoenfeld D: Chi-squared goodness-of-fit tests for the proportional hazards model. *Biometrika* 1980;67:145-153.
- SAS Institute, Inc: *SAS Language (Version 8)*, Cary, SAS Institute, 1999.
- Spitzer R, Cohen J, Fleiss J, Endicott J: Quantification of agreement in psychiatry diagnosis: a new approach. *Arch Gen Psychiatry* 1967;17:83-87.
- Shrout PE, Fleiss JL: Intraclass correlations: uses in assessing rater reliability. *Psychol Bull* 1979;2:420-428.
- McGraw KO, Wong SP: Forming inferences about some intraclass correlation coefficients. *Psychol Meth* 1996;1:30-46.
- Agresti A: *Categorical Data Analysis*. New York, Wiley, 1990.

- 30 Moher D, Jadad AR, Nichol G, et al: Assessing the quality of randomized controlled trials: an annotated bibliography of checklists. *Control Clin Trials* 1995;16:62–73.
- 31 Lionel NDW, Herxheimer A: Assessing reports of therapeutical trials. *BMJ* 1970;3: 637–640.
- 32 Badgley RF: An assessment of research methods reported in 103 scientific articles from two Canadian medical journals. *Can Med Assoc J* 1961;85:246–251.
- 33 Thomson ME, Kramer MS: Methodological standards for controlled clinical trials of early contact and maternal-infant behaviour. *Pediatrics* 1984;73:294–300.
- 34 Mosteller F, Gilbert JP, McPeck B: Reporting standards and research strategies for controlled trials: agenda for the editor. *Control Clin Trials* 1980;1:37–58.
- 35 Sacks HS, Berrier J, Reitman D, et al: Meta-analysis of randomized controlled trials. *N Engl J Med* 1987;316:450–455.
- 36 DerSimonian R, Charette J, McPeck B, et al: Reporting on methods in clinical trials. *N Engl J Med* 1982;306:1332–1337.
- 37 Peto R: Clinical trial reporting. *Lancet* 1996; 348:894–895.
- 38 Cho MK, Bero LA: Instruments for assessing the quality of drug studies published in the medical literature. *JAMA* 1994;272:101–104.