Journal of the International Neuropsychological Society (1997), **3**, 317–326.

The effect of dementia risk factors on comparative and diagnostic selective reminding norms

MARTIN SLIWINSKI,[1] HERMAN BUSCHKE,[1] WALTER F. STEWART,[2]
DAVID MASUR,[1] AND RICHARD B. LIPTON[1]

[1]The Saul R. Korey Department of Neurology, and Rose F. Kennedy Center for Mental Retardation and Human Development, Albert Einstein College of Medicine, Bronx, NY
[2]Department of Epidemiology, School of Public Health, Johns Hopkins University, Baltimore, MD

(RECEIVED December 12, 1995; REVISED July 2, 1996; ACCEPTED September 11, 1996)

Abstract

Robust comparative and diagnostic norms for the elderly are provided for the Selective Reminding Test (Buschke, 1973). Correcting for factors such as age and education level are appropriate for comparative norms, which are intended for ranking individuals with respect to their age and education matched peers. However, because age and education are both risk factors for dementia, correcting for these factors decreases test sensitivity for detecting dementia. Age- and education-corrected Selective Reminding scores have a sensitivity for detecting dementia that is 28% lower than uncorrected scores. Using information about age in combination with memory scores provided optimal discrimination of dementia. It is concluded that statistically removing the contribution of dementia risk factors from memory test scores can severely decrease discriminative validity for detecting dementia in the elderly. (*JINS*, 1997, *3*, 317–326.)

Keywords: Dementia, Selective reminding, Aging, Norms, Diagnosis, Risk factors

INTRODUCTION

Normed cognitive tests are used in elderly populations for two distinct purposes. *Comparative norms* are used to compare the performance or relative standing of an individual to a group of similar individuals using means and standard deviations, or empirically estimated percentiles. *Diagnostic norms*, usually presented in the form of cut scores, are used to optimally discriminate individuals in two or more groups, such as normal elderly *versus* demented individuals (Grober et al., 1988; Fuld et al., 1990; Masur et al., 1989, 1990, 1994). Diagnostic norms play an important clinical role in the diagnosis of dementia. As diagnostic norms are often not available, comparative norms are also used to infer impairment in individuals. Because comparative norms are designed to rank elderly individuals with respect to their age and education matched peers, they are not necessarily optimal for detecting the presence of dementia.

Comparative and diagnostic norms differ not just in application but in their nature. Comparative norms are typically "corrected" for confounding factors, such as age and education. Consequently, comparative norms are useful for answering questions such as, "Given a person's age, are they performing above average, below average or as expected?" While corrections allow us to determine how individuals compare with their peers, this procedure removes the contribution of an individual's age and educational status to diagnosis. Diagnostic norms, on the other hand, produce scores that are weighted for factors such as age that optimally discriminate individuals with dementia from those without dementia. Therefore, diagnostic norms are useful for answering questions such as, "Given a person's memory score *and* age, what is the likelihood they have dementia?"

Each type of norm relies upon a different type of statistical model that accounts for age differently. Comparative norms use statistical methods (e.g., ordinary-least-squares regression) that eliminate the effect of age on memory test scores, producing an age-corrected score. Diagnostic norms rely on methods (e.g., logistic regression) that combine memory test scores and age to form a weighted score that optimally discriminates demented from nondemented individuals. The most important difference between age-corrected and age-weighted scores is that the former contains *no* information about age (i.e., the effect of age is removed from the

Reprint requests to: Martin Sliwinski, Department of Neurology, Albert Einstein College of Medicine, 1300 Morris Park Avenue, Bronx, NY 10461. E-mail: sliwinsk@aecom.yu.edu.

score), whereas the latter combines information about memory performance *and* age to produce a score that optimally discriminates.

Though it is traditional to exclude from normative samples individuals who are in poor physical health and who meet established diagnostic criteria for dementia, many elderly individuals who do not yet meet these criteria have detectable cognitive impairment and develop diagnosable clinical dementia after several years (Masur et al., 1994; Jacobs et al., 1995). These individuals are said to be in the *preclinical* stage of dementia and including them in normative samples can bias estimates of normal performance (Morris et al., 1996; Sliwinski et al., 1996). Most normative studies fail to exclude preclinical cases, generating what we have termed *conventional* norms (Sliwinski et al., 1996). Conventional norms underestimate actual levels of cognitive function of normal elderly because individuals with preclinical dementia are included. Because individuals in the present study have been followed longitudinally, we can exclude subjects with preclinical dementia from the normative sample by using information obtained at follow-up. We use the term *robust* norms to refer to normative estimates obtained from samples that systematically exclude individuals with preclinical dementia. Although a comparison of robust and conventional norms is not the primary focus of this paper, the effect of failing to exclude preclinical cases on diagnostic sensitivity will be examined.

Comparative and Diagnostic Norms for the Selective Reminding Test

This study presents comparative and diagnostic normative data for the Selective Reminding Test (SRT) developed by Buschke (1973). This test was selected to illustrate a number of methodologic issues because of its wide use in aging and dementia research. The data were collected as part of our longitudinal studies of normal aging and dementia (Bronx Aging Study, Albert Einstein Teaching Nursing Home Program Project). Though participants were screened for dementia prior to enrollment, a large percentage of elderly participants (19%) were eventually diagnosed with dementia, creating an opportunity to develop comparative and diagnostic norms.

The present study extends previous normative work reporting normative and predictive data from the Bronx Aging Study (BAS) for the SRT in several respects (Masur et al., 1989, 1990). First, Masur and colleagues did not provide age- and education-adjusted norms for the SRT. Though they report that the association between age and education with SRT measures was not clinically significant, both of these factors have a substantial impact on expected SRT scores, and must be reflected in normative data. Second, although the ability of the SRT to predict dementia has been examined, no attempt was made to identify an optimal cut score for any of the SRT measures (Masur et al., 1989, 1990). Instead, an arbitrary cut of 2 standard deviations below the mean of nondemented elderly was used to determine sensitivity, specificity, and predictive values. Diagnostic norms should specify a cut score that optimally separates normal from demented individuals. A final limitation is that their normative sample is relatively small, consisting of only 134 persons.

In this study, we compare the utility of comparative and diagnostic norms for making diagnostic inferences regarding dementia. We hypothesized that diagnostic norms would yield more accurate diagnosis because they are designed for this purpose. We also hypothesized that comparative norms might perform better without correction for age and education. In this report, we contrast the discriminative validity of age- and education-adjusted SRT scores with the optimal classification strategies of diagnostic norms.

METHODS

Research Participants

The Bronx Aging Study initially enrolled 488 nondemented community-residing elderly persons, targeting individuals between the ages of 75 and 85 years. Participants were not eligible for the study if they had life-threatening medical conditions, severe sensory impairment (i.e., corrected visual acuity of greater than 20/200) that might interfere with study participation, or if they made more than 8 errors on the Blessed Information, Memory and Concentration Test (BIMC: Blessed et al., 1968; Fuld, 1978). Yearly assessments of each participant included a neuropsychological test battery, a neurologic examination, blood tests, and social and behavioral questionnaires. Computerized tomography (CT) and electroencephalography (EEG) were performed when subjects developed cognitive change, defined by a cumulative increase of 4 points from baseline on the BIMC, a BIMC score of more than 8 errors, or changes in behavior suggestive of dementia reported by significant others or by study personnel. A diagnosis of dementia was made according to DSM–III criteria (American Psychiatric Association, 1980) and the NINCDS-ADRDA criteria (McKhann et al., 1984) following procedures described elsewhere (Katzman et al., 1989).

We have previously demonstrated that individuals with preclinical dementia can drastically influence estimates of normal performance in the aged (Sliwinski et al., 1996). Therefore, two normative subsamples were selected from the total BAS sample according to the following criteria as outlined in Sliwinski et al. (1996):

1. The *conventional normative sample* was defined as all those individuals who qualified for inclusion in the BAS. Because not all participants received the SRT at their baseline assessment, some were first given the tests on their follow-up visits. Participants who made more than 8 errors on the BIMC at the time they received their first administration were excluded from the conventional normative sample.
2. The *robust sample* was developed to eliminate individuals with preclinical dementia at baseline, based on in-

https://doi.org/10.1017/S1355617797003172 Published online by Cambridge University Press

formation acquired by follow-up. This sample was defined using the criteria outlined above with the following additional exclusions: (1) any participant who was diagnosed as demented at any point in the course of the study was excluded; (2) any participant who was not diagnosed as demented but had fewer than 4 years of follow-up testing was excluded; and (3) any participant who was not diagnosed as demented and had a BIMC score above 8 during their first 4 years of follow-up was excluded. These criteria resulted in the selection of 236 participants out of the original 488 enrolled in the BAS.

3. A *criterion sample* of demented participants was defined for deriving diagnostic norms. This sample consisted of those participants who had an SRT administered within 1 year of the time they were diagnosed as demented. Using this criteria, 66 of the 93 participants who were eventually diagnosed as demented qualified for inclusion in the criterion sample.

Procedure

Selective reminding (Buschke, 1973) is a paradigm for multiple-trial, free-recall verbal list-learning in which participants are selectively reminded of only those items that were not recalled on the immediately preceding trial, to assess learning by recall without further presentation. On Trial 1, 12 unrelated words (Hannay & Levin, 1985) were shown and read aloud at 5-s intervals to the participant, who repeated each word aloud as it was presented. On each trial the participant attempted to recall aloud all of the words in any order. On Trials 2 to 6, the participant was reminded only of those words that were not recalled on the immediately preceding trial. The total number of words recalled over 6 trials is an index of overall learning (SUM). List items recalled on two consecutive trials without reminding are considered to have come from long-term storage (LTS), whereas items retrieved only after reminding are considered to have been retrieved from short-term memory (STR). Once an item met the criteria for LTS, all subsequent retrievals of that item are assumed to come from long-term memory (LTR). Consistent retrieval (CR) is defined as the total number of words recalled on two consecutive trials without reminding, and consistent long term retrieval (CLTR) is the number of words recalled without reminding on the last three trials (Trials 4–6). After Trial 6, 60 s of interference took place, which was followed by a final recall trial. The number of words recalled after the 60-s delay period was used to obtain a delayed recall score (DR). The procedure used for the Selective Reminding Test (SRT: Buschke, 1983) is described in detail by Masur et al. (1989).

Data Analysis

Comparative norms

Linear regression was used to measure the effect of age, education, and other demographic factors on SRT scores. Analyses for comparative norming used data from each participant's first administration of the SRT. For these analyses, education was coded to represent those with less than 6, between 7 and 9, between 10 and 12, and greater than 12 years of formal education. All regression models were examined for departures from linearity and other violations of the assumptions underlying least-squares regression.

Diagnostic norms

Logistic regression was used to determine the discriminative validity of the SRT for classifying individuals as demented. Age, education, and sex were examined to determine whether adjusting for demographic factors added to the discriminative validity of the SRT for diagnosing dementia. All significance tests reported with logistic regressions are based on the likelihood ratio. Information needed to compute predicted probabilities of dementia, as well as cut scores, is provided because classifications of those individuals falling just above or just below a cut score is made with much less confidence than the classification of individuals who fall far below or above the cut. Optimal classification derived from the logistic regression was contrasted to that obtained from a cut of -2 standard deviations using both conventional and robust comparative norms. The analysis of the SUM measure of total recall will be described in detail to illustrate several methodologic and conceptual issues.

RESULTS

Sample Demographics and Summary of SRT Measures

Demographic characteristics of the conventional, robust, and demented samples are summarized in Table 1. The bottom of Table 1 displays the means and standard deviations for the SRT measures for the robust normative sample and for the criterion dementia sample. The robust sample has higher means and smaller standard deviations for all SRT measures than the dementia sample, with the exception of STR measure, for which higher scores indicate poorer long-term recall. Correlations between age and SRT ranged between $-.18$ and $-.24$ ($p < .01$ in all cases) in the robust sample, indicating a reliable negative relationship between memory performance and age. Education also showed a strong association with the SRT measures, with correlations ranging from .22 to .31 ($p < .01$ in all cases), except for the STR measure, with which education was only correlated $-.14$ ($p = .11$). Gender was not strongly associated with any of the SRT measures, though its correlation with DR was significant at the .05 level ($r = .14$), with women recalling an average of .8 items more than men.

The dementia sample also tended to have less education and to be older at the time of their diagnosis than the robust normative sample. Logistic regression showed a significant positive association between age and the presence of dementia [odds ratio = 1.42, $\chi^2(1) = 65.5$, $p < .001$]. Education was also strongly associated with the presence of

https://doi.org/10.1017/S1355617797003172 Published online by Cambridge University Press

Table 1. Descriptive statistics for conventional, robust normal and criterion dementia samples (*SDs* in parentheses)

Variable	Conventional sample ($N = 367$)	Robust sample ($N = 236$)	Dementia sample ($N = 66$)
Age	80.2 ($SD = 3.2$)	80.0 ($SD = 3.1$)	81.7 ($SD = 3.3$)† 84.0 ($SD = 3.4$)‡
Gender			
Female	64%	63%	69%
Male	36%	37%	31%
Education (years)			
1–6	16.8%	10.7%	23.1%
7–9	33.6%	32.5%	44.6%
10–12	30.0%	32.5%	21.5%
12+	19.6%	24.3%	10.7%
Race			
White	91.6%	91.0%	90.6%
Black	8.8%	8.6%	9.4%
Other	0.6%	0.4%	
Religion			
Jewish	56.0%	55.8%	53.1%
Catholic	25.9%	25.1%	31.3%
Protestant	14.5%	14.7%	12.5%
Other	3.6%	4.3%	3.1%
Selective Reminding			
SUM	39.7 (11.1)	42.8 (8.9)	22.5 (7.5)
LTR	27.5 (14.6)	31.3 (12.7)	8.3 (8.7)
STR	12.2 (5.4)	11.5 (5.4)	14.2 (4.8)
LTS	32.0 (15.9)	36.0 (13.6)	10.8 (10.8)
CR	18.4 (11.3)	21.1 (10.1)	4.6 (5.4)
CLTR	15.3 (12.4)	17.9 (11.7)	2.2 (4.2)
DR	5.3 (3.2)	6.1 (2.8)	0.7 (1.5)

†Age at baseline SRT; ‡Age at time of diagnosis. SUM = SUM of total recall; LTR = long-term recall; STR = short-term recall; LTS = long-term storage; CR = consistent recall; CLTR = consistent long-term recall; DR = delayed recall.

dementia. Individuals with 9 or fewer years of formal education were over 2.5 times more likely to be diagnosed demented than those with at least 10 years of education [odds ratio = 2.76, $\chi^2(1) = 12.6$, $p < .01$]. Gender did not show a statistically reliable relationship with dementia.

There is a potential bias in estimating the association of age with dementia in this study. Specifically, the BAS recruited individuals between the ages of 75 and 85 years who were *not* clinically demented. Therefore, any diagnosis of dementia would have to be made at follow-up. That individuals at the time of diagnosis are an average of 4 years older than those not diagnosed is due, in part, to the fact that age is measured at baseline for the normal group and at follow-up for the demented group. Thus, the possibility exists that the effect of age in discriminating dementia is overestimated.

To examine this possibility, in a supplementary analysis a subset of normal participants (*non-cases*) were randomly selected and matched to demented subjects (*cases*) on time of follow-up. This design eliminates the bias in comparing the age of normal individuals at baseline with demented individuals at follow-up. Sample size was sufficient to allow 3:1 matching (3 non-cases matched to every 1 case). A conditional logistic regression, with subjects stratified according to length of follow-up, indicated that age was still significantly associated with the dementia [odds ratio = 1.16, $\chi^2(1) = 9.62$, $p < .01$]. Though age is still positively associated with dementia when participants are matched for length of follow-up, that association is not as strong as when age at baseline is used for nondemented individuals.

Comparative Norms for SRT Measures

Table 2 presents the regressions used for norming the SRT measures. Regression analysis was used to determine how SRT scores varied as a function of age, education, and gender. Age and education made substantial contributions to SRT scores, but gender was not associated with performance. There were no significant nonlinear effects of age for any regression. Education was initially examined using variables representing four categories: 1 to 6 years, 7 to 9 years, 10 to 12 years, and greater than 12 years of formal education. Coefficients for the lowest two categories did not differ significantly from each other, nor did the coefficients from the highest two categories differ significantly. Model fit was not altered by collapsing education into two categories: 9 or fewer and 10 or more years of education. There

https://doi.org/10.1017/S1355617797003172 Published online by Cambridge University Press

Table 2. Robust sample regressions for SRT measures

SRT Measure	Variable	Coefficient	SE	Multiple *R*	Residual MSE
SUM	Intercept	92.12	14.27		
	Age	−.64	.18	.30	8.46
	Educ. ≥10	4.12	1.11		
DR	Intercept	19.25	4.58		
	Age	−.17	.06	.26	2.71
	Educ. ≥10	1.17	.36		
LTS	Intercept	110.05	22.06		
	Age	−.96	.28	.28	13.08
	Educ. ≥10	5.34	1.72		
LTR	Intercept	109.68	20.54		
	Age	−1.01	.26	.30	12.18
	Educ. ≥10	5.11	1.60		
STR	Intercept	−17.56	8.91		
	Age	.37	.11	.22	5.28
	Educ. ≥10	−.99	.69		
CLTR	Intercept	95.24	18.81		
	Age	−.99	.24	.30	11.15
	Educ. ≥10	4.44	1.47		
CR	Intercept	83.73	16.23		
	Age	−.81	.20	.32	9.63
	Educ. ≥10	4.28	1.27		

The coefficient for education refers to the adjustment in the estimated SRT measure for 10 or more years of education. PPV = positive predictive value; NPV = negative predictive value; Educ. = education; SE = standard error; residual MSE = residual mean squared error (standard deviation of residuals).

were no nonlinear effects of age in any of the regressions, and none of the regressions violated the assumption of homoscedasticity. Residuals for the DR and CLTR measures were shown to deviate significantly from normality [$z = 2.1$, $p < .05$, and $z = 3.5$, $p < .01$, respectively]. However, there was not a large discrepancy between empirically estimated percentiles and those computed from the sample mean for either of these two measures. Therefore, it was deemed acceptable to use the information provided in Table 2 to compute age- and education-corrected scores for each of the SRT measures.

Diagnostic Norms for SRT Measures

A series of logistic regressions was performed to determine how accurately the various SRT measures could classify participants who were diagnosed as demented within 1 year of testing. This time interval was selected to approximate the clinical circumstance where concurrent diagnosis is the objective. Including age improved the discrimination in all cases, but adding education and gender failed to improve classification for any of the measures. Adding a term for age squared did not improve predictive validity. The analysis of the SUM measure of total recall will be described in detail to illustrate several methodologic and conceptual issues.

Using baseline age for the normal group clearly overestimates the association between age and dementia. The analysis in which normals and dements are matched on length of follow-up eliminates this systematic overestimation. However, the results from the matched analyses cannot be used for obtaining predicted probabilities and cut scores for dementia. Matched analyses can provide such probabilities only for individuals who can be placed within a stratum of the variable on which the matching was done. Since the variable (length of follow-up) used for matching is not applicable in clinical settings, the matched analyses are useful only for estimating the effect of age, and not for providing cut scores appropriate for clinical diagnosis.

One solution to this problem is to estimate the logistic regression for the SRT measures and age after normals and dements are matched for length of follow-up, as reported above. Then, use baseline SRT and age for the normals to fit a generalized linear model (McCullagh & Nelder, 1989) that estimates the coefficient for the SRT measure while holding the coefficient for age fixed to the estimate from the matched analysis. This remedies the overestimation of the coefficient for age that results from comparing age at baseline for normals to age at follow-up for dements, and it generates a model that provides information useful for establishing cut scores and predicted probabilities of dementia. Table 3 displays the results of this analysis, including regression coefficients, odds ratios, sensitivity, specificity, and positive and negative predictive value for all the SRT measures.

The odds ratio for the SUM score (top of Table 3) indicates that for every point decrease, the probability of dementia increases by a multiplicative factor of 1.31. Similarly, for every year older a participant is, the likelihood of de-

https://doi.org/10.1017/S1355617797003172 Published online by Cambridge University Press

Table 3. Robust sample logistic regressions (GLIMs) for age and SRT measures predicting to criterion sample of clinically demented elderly

Variable	Coefficient	SE	Odds ratio	Sens	Spec	PPV	NPV
Intercept	−4.77	1.08					
SUM	−.27	0.04	1.31	78.8	96.6	86.7	94.2
Intercept	−10.96	0.28					
DR	−.85	0.12	2.35	81.8	92.8	76.1	94.8
Intercept	−10.16	0.38					
LTR	−.18	0.02	1.19	72.7	93.6	76.2	92.5
Intercept	−14.22	0.39					
STR	.08	0.03	1.08	6.1	98.3	50.0	78.9
Intercept	−10.18	0.38					
LTS	−.14	0.02	1.15	74.2	93.64	76.6	92.9
Intercept	−11.41	0.24					
CLTR	−.26	0.04	1.30	74.2	91.9	72.1	92.7
Intercept	−10.59	0.33					
CR	−.18	0.04	1.19	71.2	92.4	72.3	91.9

Odds ratios reflect how the odds of dementia changes for every 1-point decrease in SRT performance. The offset for age (.147 × age) was calculated using coefficient for age from the conditional logistic regression in which normals were matched to dements on length of follow-up. SE = standard error; Sens = sensitivity; Spec = specificity; PPV = positive predictive value; NPV = negative predictive value.

mentia increases by a factor of 1.16. Figure 1 presents predicted probabilities of dementia as a function of SRT SUM score for ages 75, 80 and 85 years as derived from the logistic regression described in Table 3. A score of 25 corresponds to a probability of only .38 of dementia for 75-year-olds, but the same score indicates a probability of .56 for 80-year-olds. And for 85-year-olds, a score of 25 on the SUM corresponds to a .73 probability of dementia.

Comparative *Versus* Diagnostic Norms for Classification of Dementia

In the absence of validated diagnostic cut scores, it is common practice to take deviant scores (i.e., < 2 *SD*s below the mean) as indicative of impairment. A cut of −2 standard deviations was used to classify individual subjects as demented using two different SRT SUM scores: (1) the con-

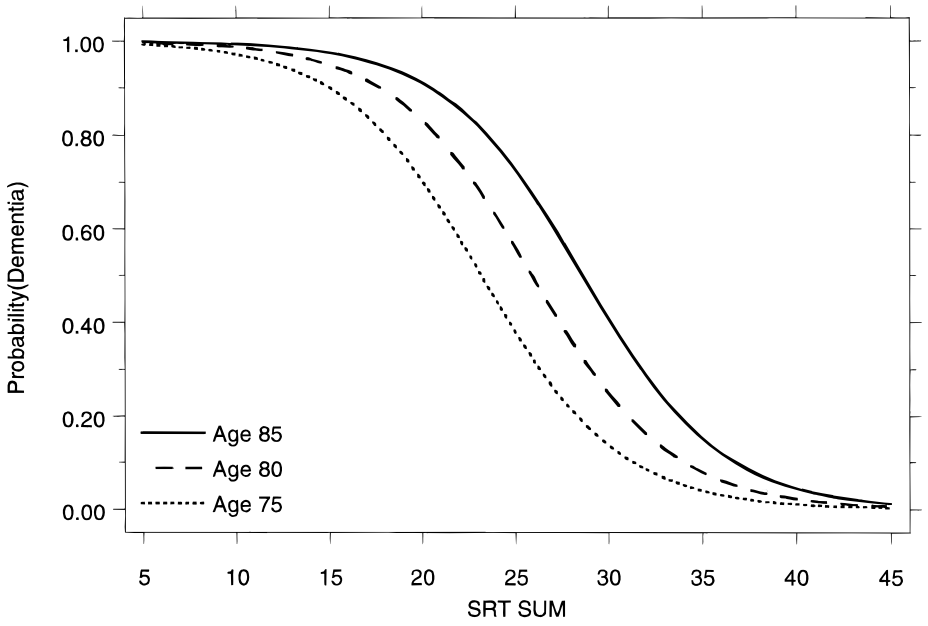

Fig. 1. Predicted probabilities of dementia as function of SRT SUM scores for 75-, 80-, and 85-year-olds are shown. The functions indicate that the probability of dementia for a given SUM score increases with age.

https://doi.org/10.1017/S1355617797003172 Published online by Cambridge University Press

ventional age-corrected score, (2) the robust age-corrected score. To examine how well this strategy fares compared to optimal cut scores derived through logistic regression, classification accuracies were compared. The conventional age-corrected SUM score was by far the worst in terms of discriminative validity, having a sensitivity of only 34.9% (and a specificity of 98.7%). The robust age- and education-corrected norms captured more dements than did the conventional norms with a small decrease in specificity (sensitivity = 50.0%, specificity = 96.7%); however this still did not approach the discrimination obtained by the logistic regression, which used both SUM and age for classification (sensitivity = 78.7%, specificity = 96.1%).

There are two distinct effects operating that need to be disentangled, namely including cases with preclinical dementia and correcting for age. It is possible that either of these affects test discrimination or the optimal impairment cut score, or both. To address this issue, receiver operating characteristic (ROC) curves were plotted and compared (see Figure 2). ROC curves show the tradeoff between sensitivity and specificity: the closer the curves run to the upper left corner of the plot, the higher the sensitivity for a given level of specificity. Comparison of the ROC curves shows little difference between conventional and robust age-corrected SUM scores for discriminating dementia. Neither of these scores performs as well as the SUM score and age in combination. This confirms that including preclinical cases in the normative sample affects the optimal impairment cut score, but does not reduce test discrimination. However, correcting for age *does* reduce the discriminative ability of the SUM. This was confirmed by using logistic regression to derive an optimal classification cut using the age-corrected SRT scores. Using the optimal cut score as determined by logistic regression, age- and education-corrected memory scores had good specificity (96.1%), but the sensitivity was only 52.3%. This final result indicates that even using an empirically determined optimal cut, age-corrected scores are much less sensitive in detecting dementia than are age-weighted scores.

DISCUSSION

This research demonstrates that using age- and education-corrected SRT scores, while appropriate for comparative norming, results in substantially reduced predictive validity for classifying individuals as demented. Using memory performance and age in combination yielded the best discrimination of dementia for several of the SRT measures.

Age Adjustments in Comparative and Diagnostic Norms

Because performance on cognitive tests, especially those tapping memory, declines with advancing age, comparative norms become more forgiving with advancing age. That is, a given raw score on a memory test may indicate relatively better performance for an 85-year-old than for a 75-year-old. For example, assume that the average score on a memory test is 45 for 75-year-olds and is 40 for 85-year-olds, with a standard deviation of 5. Now, assume that a 75- and an 85-year-old each have a score of 35 on the test. The age-corrected standardized *z* scores are −2 for the 75-year-old and −1 for the 85-year-old. Thus, the same raw score (35) corresponds to approximately the 2nd percentile for the 75-year-old and to the 16th percentile for the 85-year-old. This kind of age correction makes intuitive sense: since the

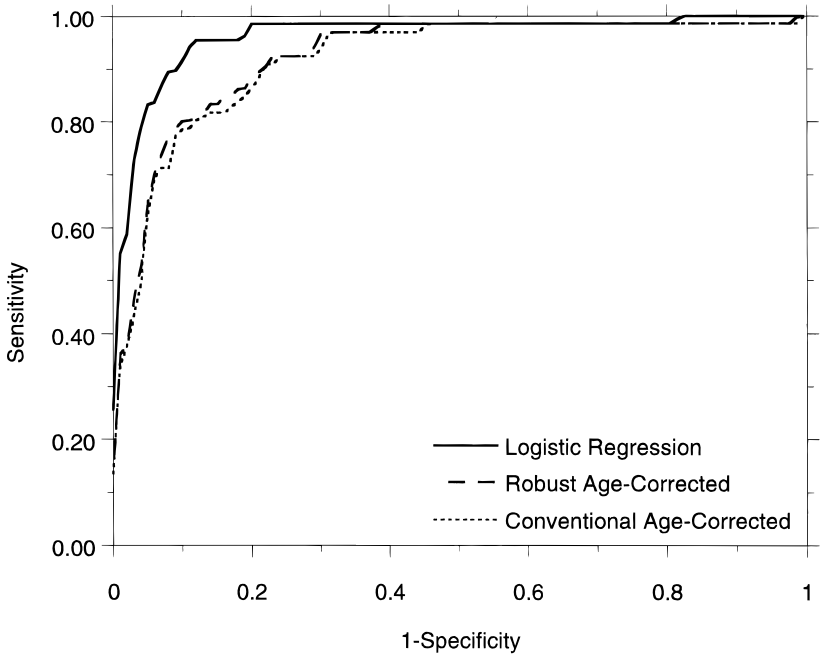

Fig. 2. The ROC curves for age-corrected SRT SUM scores from the conventional and robust normative samples, and for the optimal combination of SUM and age (logistic regression). There is little difference in discrimination between conventional and robust age-corrected scores, as seen by their overlapping ROC curves. Age-corrected scores discriminate dementia much less effectively than the optimal combination of SUM and age.

https://doi.org/10.1017/S1355617797003172 Published online by Cambridge University Press

average memory score declines as a function of age, a given memory score should signify relatively better performance for an older person than for a younger person.

If performance on this hypothetical memory test is used to infer memory impairment, an arbitrary cut of −2 standard deviations might be adopted. That is, if an individual's age adjusted score is 2 or more standard deviations below the mean, then that is taken as evidence of memory impairment and dementia. This cut score translates to a raw score of 35 for 75-year-olds and 30 for 85-year-olds on our hypothetical memory test. Therefore, a score of 32, for example, would be considered impaired for a 75-year-old, but not for an 85-year-old. At first glance, this type of age correction also seems intuitive: an age-corrected score that is more deviant (i.e., further from the age-adjusted mean) is more likely to indicate dementia than an age-corrected score that is less deviant. However, this thinking ignores the dramatic increase in the baserate of dementia with age. The prevalence of dementia doubles approximately every 5 years after the age of 65, and longitudinal studies have demonstrated that the incidence of dementia continues to rise through the age of 89 (Katzman, 1993; Katzman & Kawas, 1994).

Given the strength of the association between age and dementia, knowing nothing but a person's age provides considerable information regarding the likelihood that the person has a dementing illness. Using a person's age and memory score in combination might better discriminate between normal aged and those with dementia than using a memory score that is corrected for age. In this case, a given memory score could signify a higher probability of dementia for an older person compared to a younger person. That is, instead of making impairment cuts more stringent (i.e., lower) for older persons, using the information that age contains about the baserate of dementia might indicate that optimal impairment cuts are constant across age, or even that they become more lenient (i.e., higher) in older persons.

Age-Corrected Comparative Norms are Inappropriate for Diagnosing Dementia

Comparative norms provide standardized scores, so that an individual's performance can be judged relative to the performance of other, similar individuals. Age corrections account for the influence of age on the cognitive measure and are essential for tests tapping cognitive processes such as memory, because performance changes significantly as a function of age. However, correcting for age in this manner dramatically reduces the sensitivity of SRT measures for detecting dementia. Using the corrected scores from the conventional sample resulted in a sensitivity approximately 43% lower than for the uncorrected raw scores. Using robust samples helps, but does not eliminate this problem: the sensitivity of the corrected SUM scores derived from the robust sample is 28% lower than for the uncorrected raw SUM scores.

These results strongly suggest that comparative norms are suboptimal for diagnosing dementia in elderly individuals. In the case of conventional comparative norms, the contamination by individuals with preclinical dementia coupled with corrections for age artificially diminished test sensitivity of the SRT to a point where many individuals with dementia are missed. Examination of the ROC curves (Figure 2) indicates that contamination by preclinical dementia reduces test sensitivity by causing a lower than optimal cut score for classification to be selected. However, the ROC curves show that age- and education-corrected scores are not less sensitive than uncorrected scores simply because of the selected cut score. That is, there was no cut score (e.g., −1 *SD*, −1.5 *SD*, etc.) for the age-corrected norms that approached the level of discrimination obtained by the age-weighted scores from logistic regression. Therefore, we conclude that correcting for age and education reduces test sensitivity by decreasing the discriminative validity of the memory test.

Using Memory Scores and Age in Combination Improves Identification of Dementia

In no case did age-corrected scores approach the level of discrimination obtained when using memory scores and age in combination. Even memory scores without age correction performed better than memory scores with age correction. These results demonstrate the need for independently establishing diagnostic norms and comparative norms, since the two serve fundamentally different functions. Although it is not difficult to compute predicted probabilities of dementia based on the information provided in Table 3, the Appendix displays these probabilities for the SUM score for ages 75, 80, and 85 years.

We argued that using age-corrected scores for diagnosis will result in higher (i.e., more stringent) cut scores for younger individuals and lower (i.e., more lenient) cut scores for older individuals. It was further argued that these adjustments will diminish the discriminative validity of cognitive tests for detecting dementia because age is *positively* associated with the presence of dementia. The present findings support this argument. The age corrections require an implicit assumption that there is no association, or a *negative* association between age and dementia. Although the effects of age on optimal impairment cut scores must be determined empirically, on a test-by-test basis, it is highly improbable that making impairment cut scores higher for older individuals (as is the case when using age-corrected scores) would result in an optimal discrimination between normal and demented elderly.

Limitations of the Present Study

Age range

The age range of the normal elderly is relatively narrow, with a minimum age of 74 and a maximum age of 88 years.

https://doi.org/10.1017/S1355617797003172 Published online by Cambridge University Press

Applying the regression equations to obtain age- and education-adjusted standard scores or probabilities of dementia should be restricted to individuals falling within this age range. The present results provide no information regarding the validity of using these normative regressions in individuals whose age falls outside this range.

Age adjustments

The manner in which the BAS sample was obtained ensured that individuals could be diagnosed only during follow-up visits. This caused the age difference between individuals at the time of diagnosis and normal individuals at baseline to be artificially large. Analyses in which dements and normals were matched on length of follow-up confirmed the association between dementia and age, but also showed that using baseline data from normals overestimates the effect of age. The problem of adjusting for age was addressed by fitting a generalized linear model using baseline SRT and age for the normals while holding the coefficient for age fixed to the value obtained from the matched analysis. Although this strategy eliminates the systematic overestimation of the association between age and dementia caused by using baseline data for the normals and follow-up data for the dements, it does not guarantee that this estimate of the coefficient for age will generalize to other settings.

Baserates

The baserate of clinically diagnosed dementia is 21.8% (66/302) in the sample used for norming the SRT. The optimal cut scores, and estimates of positive and negative predictive value depend heavily upon the baserate. If the baserate is dramatically different, the optimal cut score for diagnosing dementia will also be different. The robust comparative norms do not depend upon the baserate of dementia. Using the SRT in a clinical setting where the baserate of dementia may be closer to 50% would require more lenient cut scores to maintain the predictive values reported in the present study.

Generalizability

The samples used for norming cannot be considered a random sample of elderly individuals living the community. We evaluated a sample of volunteers who passed a mental status exam at baseline and remained actively engaged in the study for at least 4 years. This subset of elderly is very different from the general population of elderly individuals residing in the community. However, the alternative of not screening and following elderly individuals to verify their cognitive status is also undesirable. We expect that a random sample of community-based elderly could contain a high proportion of cases with preclinical dementia.

Because comparisons were between two well-defined groups (robust normals *vs.* clinical dementia), caution should be used when interpreting the discrimination indices for the SRT measures reported in Table 3. Consequently, the reported sensitivity and specificity will likely overestimate the classification accuracy of the SRT in clinical or research settings in which the nondemented and demented individuals are less well defined.

CONCLUSIONS

The present findings point to important methodologic and conceptual issues that deserve special attention when norming cognitive tests in the elderly. First, using samples contaminated by preclinical dementia will result in the selection of less-than-optimal impairment cut scores for detecting dementia. Second, and most important, norms devised for comparative purposes perform poorly when used for classifying individuals as demented because correcting for age and education significantly reduces the predictive validity of all SRT measures for discriminating dementia. Age and education corrections are appropriate only when trying to estimate mean recall on the SRT for the purposes of comparing one individual's performance to a population of similarly aged and educated individuals. Further research is needed to demonstrate how adjusting for age affects the discriminative properties of other cognitive tests used for identifying dementia in elderly populations. Considering the present findings, researchers and clinicians should be extremely cautious about using age-corrected memory scores for detecting dementia-related memory impairment.

ACKNOWLEDGMENTS

Supported by National Institute on Aging Grants R29 AG12448, AGO3949 (Project 2) and National Institute of Child Health and Human Development Grant HD-01799.

REFERENCES

American Psychiatric Association. (1980). *Diagnostic and statistical manual of mental disorders* (3rd ed.). Washington, DC: Author.

Blessed, G., Tomlinson, B.E., & Roth, M. (1968). The association between quantitative measure of dementia and of senile changes in the cerebral grey matter of elderly subjects. *British Journal of Psychiatry*, *114*, 797–811.

Bondi, M.W., Monsch, A.U., Galasko, D., Butters, N., Salmon, D.P., & Delis, D.C. (1994). Preclinical markers of dementia of the Alzheimer Type. *Neuropsychology*, *8*, 374–384.

Buschke, H. (1973). Selective reminding for analysis of memory and learning. *Journal of Verbal Learning and Verbal Behavior*, *12*, 543–550.

Folstein, M.F., Folstein, S.E., & McHugh, P.R. (1975). "Mini-mental state." A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, *12*, 189–198.

Fuld, P.A. (1978). Psychological testing in the differential diagnosis of the dementias. In R. Katzman, R.D. Terry, & K.L. Bick (Eds.), *Alzheimer's disease: Senile dementia and related disorders* (pp. 185–193). New York: Raven Press.

Hannay, J.H. & Levin, H.S. (1985). Selective reminding test: An examination of the equivalence of four forms. *Journal of Clinical and Experimental Neuropsychology*, *7*, 251–263.

Jacobs, D.M., Sano, M., Dooneief, G., Marder, K., Bell, K.L., & Stern, Y. (1995). Neuropsychological detection and characterization of preclinical Alzheimer's disease. *Neurology*, *45*, 957–962.

Katzman, R. (1993). Education and the prevalence of dementia in Alzheimer's disease. *Neurology*, *43*, 13–20.

Katzman, R., Aronson, M., & Fuld, P.A. (1989). Development of dementing illness in an 80-year-old volunteer cohort. *Annals of Neurology*, *25*, 317–324.

Katzman, R. & Kawas, C. (1994). The epidemiology of dementia and Alzheimer's disease. In R. Katzman, R.D. Terry, & K.L. Bick (Eds.), *Alzheimer's disease* (pp. 105–122). New York: Raven Press.

La Rue, A. & Jarvik, L. (1987). Cognitive function and prediction of dementia in old age. *International Journal of Aging and Human Development*, *25*, 79–88.

Masur, D.M., Fuld, P.A., Blau, A.D., Crystal, H., & Aronson, M. (1990). Predicting development of dementia in the elderly with the selective reminding test. *Journal of Clinical and Experimental Neuropsychology*, *12*, 529–538.

Masur, D.M., Fuld, P.A., Blau, A.D., Thal, L.J., Levin, H.S., & Aronson, M. (1989). Distinguishing normal and demented elderly with the selective reminding test. *Journal of Clinical and Experimental Neuropsychology*, *11*, 615–630.

Masur, D.M., Sliwinski, M., Lipton, R.B., Blau, A.D., & Crystal, H.A. (1994). Neuropsychological prediction of dementia and the absence of dementia in healthy elderly. *Neurology*, *44*, 1427–1432.

McCullagh, P. & Nelder, J.A. (1989). *Generalized linear models*. (2nd ed.). New York: Chapman & Hall.

McKhann, G., Drachman, D., Folstein, M., Katzman, R, Price, D., & Stadlan, E. (1984). Clinical diagnosis of Alzheimer's disease: Report of the NINCDS-ADRDA Work Group under the auspices of the Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurology*, *34*, 939–944.

Morris, J.C., Storandt, M., McKeel Jr., D.W., Rubin, E.H., Price, J.L., Grant, E.A., & Berg, L. (1996). Cerebral amyloid deposition and diffuse plaques in "normal" aging: Evidence for presymptomatic and very mild Alzheimer's disease. *Neurology*, *46*, 707–719.

Sliwinski, M., Lipton, R.B., Buschke, H., & Stewart, W.F. (1996). The effect of pre-clinical dementia on estimates of normal cognitive function in aging. *Journal of Gerontology: Psychological Sciences*, *51B*, P217–P225.

Appendix

Probability of dementia as a function of age at testing and SUM of total recall

SUM score	75	80	85
>35	<.04	<.07	<.13
35	.04	.08	.15
34	.05	.10	.19
33	.07	.13	.23
32	.08	.16	.29
31	.11	.20	.34
30	.14	.25	.41
29	.17	.30	.47
28	.21	.36	.54
27	.26	.42	.61
26	.32	.49	.67
25	.38	.56	.73
24	.44	.62	.78
23	.51	.68	.82
22	.58	.74	.86
21	.64	.79	.89
20	.70	.83	.91
<20	>.75	>.85	>.93

https://doi.org/10.1017/S1355617797003172 Published online by Cambridge University Press