Computer-Automated Dementia Screening Using a Touch-Tone Telephone

James C. Mundt, PhD; Kae L. Ferber, MD; Matthew Rizzo, MD; John H. Greist, MD

**Background:** This study investigated the sensitivity and specificity of a computer-automated telephone system to evaluate cognitive impairment in elderly callers to identify signs of early dementia.

**Methods:** The Clinical Dementia Rating Scale was used to assess 155 subjects aged 56 to 93 years (n=74, 27, 42, and 12, with a Clinical Dementia Rating Scale score of 0, 0.5, 1, and 2, respectively). These subjects performed a battery of tests administered by an interactive voice response system using standard Touch-Tone telephones. Seventy-four collateral informants also completed an interactive voice response version of the Symptoms of Dementia Screener.

**Results:** Sixteen cognitively impaired subjects were unable to complete the telephone call. Performances on 6 of 8 tasks were significantly influenced by Clinical Dementia Rating Scale status. The mean (SD) call length was 12 minutes 27 seconds (2 minutes 32 seconds). A subsample (n=116) was analyzed using machine-learning methods, producing a scoring algorithm that combined performances across 4 tasks. Results indicated a potential sensitivity of 82.0% and specificity of 85.5%. The scoring model generalized to a validation subsample (n=39), producing 85.0% sensitivity and 78.9% specificity. The κ agreement between predicted and actual group membership was 0.64 (P<.001). Of the 16 subjects unable to complete the call, 11 provided sufficient information to permit us to classify them as impaired. Standard scoring of the interactive voice response–administered Symptoms of Dementia Screener (completed by informants) produced a screening sensitivity of 63.5% and 100% specificity. A lower criterion found a 90.4% sensitivity, without lowering specificity.

**Conclusions:** Computer-automated telephone screening for early dementia using either informant or direct assessment is feasible. Such systems could provide wide-scale, cost-effective screening, education, and referral services to patients and caregivers.

Arch Intern Med. 2001;161:2481-2487

Alzheimer disease (AD) is the most common cause of dementia for elderly patients. Years may pass following symptomatic onset before diagnosis,1 and current treatments may slow but will not reverse the progressive cognitive decline.2 Earlier detection and recognition of dementia would permit more effective use of available treatments, better opportunity to educate patients and families, and time to develop social support systems and implement important financial and legal plans.3 Key to early detection and recognition of AD are effective systems for patient screening. Screening approaches to identify cognitive impairment in the elderly have included direct patient evaluation4,5 and collateral informant questionnaires.6-9 Both approaches are effective for accurately identifying unrecognized dementia in patients. However, large community screening efforts to identify persons for diagnostic evaluation are time consuming and resource intensive.

Interactive voice response (IVR) systems integrate telecommunications networks with computer-automated processing. Programs using IVR systems have become commonplace in society for automated call routing and access to banking records, airline schedules, and local theater listings. Health care delivery and monitoring systems have increasingly used IVR systems across a range of problems from psychiatric and behavioral disorders (such as depression, anxiety, obsessive-compulsive disorder, and substance abuse10-13) to hypertension monitoring.14 Such systems have been successfully implemented to monitor the functional status of community-residing elders enrolled in home care programs.15

From Healthcare Technology Systems Inc, Madison, Wis (Drs Mundt and Greist); the Department of Internal Medicine, Geriatrics Section, Dean Medical Center, Madison (Dr Ferber); and the Department of Neurology, University of Iowa, Iowa City (Dr Rizzo). Dr Ferber is currently with Anchor Health Centers, Naples, Fla.

©2001 American Medical Association. All rights reserved.
PARTICIPANTS AND METHODS

PARTICIPANTS

For our study, 155 subjects were recruited from both a geriatrics practice affiliated with the Dean Medical Center, Madison, Wis (n=91), and from ongoing research at the Department of Neurology, University of Iowa, Iowa City (n=64). Geriatric patients with normal cognition and/or a diagnosis of mild dementia who scheduled appointments with the Dean Medical Center were invited to participate. Research subjects from the University of Iowa were recruited from a memory disorders clinic and from a federally funded mobility study of elderly licensed drivers. Informed consent was obtained from all participants in accordance with required federal and institutional guidelines. Study participants were not compensated for participation. Subjects ranged in age from 56 to 93 years (mean [SD], 76.7 [7.0] years) with 6 to 22 years of education (mean [SD], 13.3 [3.0] years; 13.5% had not graduated high school, 51.0% graduated high school, 22.6% had 2- or 4-year college degrees, and 12.9% had earned graduate degrees). The sample included 98 women and 57 men; 61.3% were married, 31.0% widowed, 5.2% divorced, and 2.6% never married. In addition to the 155 subjects participating directly in the testing procedures, collateral informants accompanied 74 subjects to their appointment (46 spouses, 27 children or grandchildren, and 1 other). During a separate, independent telephone call to the system, these informants completed an IVR-administered Symptoms of Dementia Screener (SDS).9 The SDS is an 11-item checklist of dementia symptoms often noted by family members and caregivers prior to detection, evaluation, and diagnosis by medical staff. Previous research9 using telephone interviews suggested that positive endorsement of 5 or more symptoms by an informant is associated with risk of dementia. This research is the first attempt to apply this screener with IVR technology.

Each subject was given a Mini-Mental State Examination (MMSE),4 and a trained clinician provided ratings for the Clinical Dementia Rating Scale (CDRS).21,22 The CDRS obtains an impairment rating for each of 6 functional areas: memory, orientation, judgment and problem solving, community affairs, home and hobbies, and personal care. Clinical ratings in each area are anchored to explicit descriptions of patient symptoms and functional difficulties, resulting in impairment rating values of 0 (none), 0.5 (questionable), 1 (mild), 2 (moderate), or 3 (severe). Scoring for the CDRS considers impairment ratings across all 6 areas using memory impairment ratings as the primary index and ratings of impairment in the other domains as secondary indexes.22 The resulting CDRS scores are used to stage dementia levels. A score of 0 indicates no cognitive impairment; 0.5, uncertain or deferred diagnosis; 1, mild stage of dementia; 2, moderate stage of dementia, and 3 to 5, profound or terminal dementia. In the present study, 74 (47.7%) of the subjects had a CDRS score of 0; 27 (17.4%), a score of 0.5; 42 (27.1%), a score of 1; and 12 (7.7%), a score of 2. None of the study participants had a CDRS score of 3 to 5.

APPARATUS

The screening procedures were designed and programmed using a Conversant MAP40 IVR server (Lucent Technologies, Murray Hill, NJ) maintained by Healthcare Technology Systems, Inc (Madison, Wis). The tasks were programmed as separable testing modules, and instructions were provided by the IVR system prior to each task. All responses were collected using standard Touch-Tone telephones. Evaluative feedback regarding task performance (ie, correct or incorrect) was not provided.

TEST MODULES

- Subjective Memory Complaint. Subjects were asked if they often had difficulty remembering names of family or friends, finding words or where objects had been left, or used notes to avoid forgetting. Subjects giving positive responses were asked to rate the severity of problems such difficulties caused (none, small, moderate, serious).
- Orientation. Subjects responded to 5 questions pertaining to their orientation in time. They were asked to enter (1) the 4 digits of the current year; (2) the current season; (3) the current month; (4) the current day of the month; and (5) the current day of the week. Responses

RESULTS

Including reference to the Table, Figure 1 shows the data collection procedures and analysis plan. The figure accurately depicts data collection from study subjects and informants as 2 separate calls at different times from different locations. Both calls were placed to the same telephone number and processed by one IVR system (ID was used to identify the appropriate IVR script to apply).

IVR TASK PERFORMANCE

Of 155 subjects participating in this study, 16 were unable to complete the call and “hung up” on the system.
were scored as correct or incorrect and totaled to produce a score of 0 to 5.

- **Alphabetic Translation.** Subjects were asked to use the letters printed on the telephone keys to spell the word “FUN” (3-8-6) and given the context as in “The party was fun to assure comprehension. A score of 0 to 3 reflected the number of correctly sequenced key presses.

- **Immediate Recall.** Subjects heard the digit sequence “2-7-6-0-4” and were asked to enter these digits in the same order. This procedure was repeated 3 times. Each trial was scored 0 to 5 based on the number of correctly sequenced key presses and then totaled to produce a score of 0 to 15.

- **Directed Key Pressing.** Subjects were directed to press particular keys a specific number of times (eg, “Press the ‘7’ key 3 times” “Press 6 times on the ‘3’ key”). Subsequent stimuli were presented after a 2-second delay without a key press. Performance continued until 30 seconds had elapsed since the start of the task. Each series of key presses was scored as correct or incorrect and totaled to produce a score of 0 to 5.

- **Delayed Recall.** Following the directed key-press task, the subjects were asked to recall the 5-digit sequence of the immediate recall trials. The score (0-5) reflected the number of correctly ordered key presses.

- **Auditory Spatial Relations.** Subjects heard an auditory description of key locations (1-9) according to the standard 3×3 matrix on most telephones (top row, 1-3; middle row, 4-6; and third row, 7-9) and were asked to press the identified key. For example, the “top-left” key corresponds to the “1” key; the “right-bottom” key would be the “9” key. Presentation of the next descriptor was prompted by any key press or proceeded after a 3-second delay without a response. Total task duration was 30 seconds, and the score (0-9) reflected the number of correct key presses made.

- **Backward Digit Span.** On 3 successive trials, subjects heard a 4-digit sequence of numbers (different sequence each trial) and asked to press the identified keys in reverse order. Trials were scored with respect to the number of correctly sequenced key presses (0-4) and were then totaled to produce a score of 0 to 12.

- **Semantic Comprehension.** Subjects heard 6 declarative statements and were asked to judge whether each statement made sense or not. Three statements made sense (eg, “The woman burned herself badly when she spilled a pot of hot soup on herself while preparing dinner”), and 3 did not (eg, “We wanted to cut down a tree in our front yard, so we went to the garage to get our hammers”). The mean (SD) length of the statements was 24 (3.9) words (range, 18-30) with a Flesch-Kincaid grade level of 6.5 (range, 3.7-8.6). Responses were scored for accuracy and were totaled to a score of 0 to 6.

**PROCEDURES**

Subjects were provided with a Touch-Tone telephone and quiet space from which to make the call from the study site clinics. Study staff provided each subject with a unique 4-digit identification number to enter at the beginning of the call as well as the toll-free number to dial. Research staff provided verbal assistance in dialing the number and entering the identification number (ID) if needed, but no further assistance in responding to the IVR system was provided. All instructions for completing the IVR tasks were provided by the IVR system at the time the data were collected. After call completion, subjects rated the overall difficulty of the testing procedures and each specific task on a 1 to 5 scale (very easy to very difficult). Feedback was also obtained regarding the clarity of the instructions and whether the task requirements were understood. Paper-based forms labeled by the ID of the subject contained the demographic information, MMSE and CDRS scores, collateral informant ID, and patient feedback and were forwarded to Healthcare Technology Systems Inc for integration with the IVR performance data.

When collateral informants were available to participate, they were removed from the vicinity of the subjects while the testing call was completed. They were not permitted to provide assistance, nor were they given knowledge of the subjects’ performance. The collateral informants completed the SDS during a separate telephone call outside the presence of the subject. Each collateral informant was given a unique ID that allowed the IVR to branch to the SDS delivery module and allowed the data collected to be linked to the ID of the target subject.

These noncompleters ranged in age from 58 to 88 years with a mean MMSE score of 21.4 (range, 11-29). Five were moderately demented, 9 were mildly demented, and 2 had questionable diagnoses based on their CDRS scores. The median time to hang-up was 8 minutes 40 seconds. All subjects without dementia (CDRS score, 0; n=74) were able to complete the call.

The Table gives the mean call length and task performance scores for the 139 subjects completing the IVR call, stratified by CDRS score. The mean (SD) time to complete the call was 12 minutes 27 seconds (2 minutes 32 seconds); subjects with greater cognitive impairment required more time to complete the call. Task performances also reflect the clinicians’ ratings of impairment, evidenced by significant analyses of variance on all tasks except the delayed recall (floor effect) and directed keypress (ceiling effect) tasks. Generally, post hoc comparisons found that the performances of subjects with mild to moderate dementia differed significantly from subjects without dementia; the intermediate performances of subjects with uncertain diagnoses (CDRS score, 0.5) did not differ from one or both of these subject groups using protected Newman-Keuls comparisons.

The bottom of Figure 1 shows how machine-learning methods were used to investigate whether combining performances across tasks could differentiate unimpaired subjects (CDRS score, 0) from cognitively impaired subjects (CDRS score, ≥0.5). Subjects were dichotomized and randomly assigned to a model “development” sample (P = .75) or a “validation” sample (P = .25). Random assignment was examined by comparing the mean age, education, and MMSE, CDRS, and total IVR task scores between the samples. No statistically significant differences were found.

Data from the development sample were analyzed using QUEST (Quick, Unbiased, Efficient Statistical
Five of 8 terminal nodes resulted in a cognitively impaired classification. Unless 1 of these circumstances was met, subjects were classified as unimpaired. This model correctly classified 50 of 61 subjects with questionable, mild, or moderate dementia (82.0% sensitivity) and 47 of 55 subjects with a CDRS score of 0 (85.5% specificity) in the development sample. Positive and negative predictive values were 0.862 and 0.810, respectively. Such methods, however, use ad hoc statistical properties of the sample and clinical judgments to generate the decisional models. Independent validation is needed to evaluate generalizability.

The decision rules of Figure 2 were applied to the 39 subjects held out of the model development analysis. Of 20 subjects with questionable, mild, or moderate dementia, 17 were predicted to be cognitively impaired (85.0% sensitivity); 15 of 19 subjects with a CDRS score of 0 were predicted not to be impaired (78.9% specificity). Prospective positive and negative predictive values were 0.810 and 0.833. The \( \kappa \) coefficient of agreement between the predicted and true group membership was 0.64 (\( P < .001 \)).

The classification tree used call noncompletion to classify subjects as cognitively impaired (100% specificity). However, in 13 of 16 hang-ups, 1 or more of the tasks were completed, of which 8 had an orientation score of 3 or less; 2 produced scores less than 3 on the “spell FUN” task; and 1 provided immediate recall data that would result in classification as impaired, regardless of sentence comprehension performance. Thus, of 16 subjects not completing the call, 11 (69%) provided sufficient data to warrant classification as impaired before hanging up.

In summary, when the derived scoring algorithm was applied to the complete study sample of 155 patients, 62 (84%) of the 74 patients with a CDRS score of 0 were classified as unimpaired. A positive screening result for cognitive impairment was found in 17 (63%) of the 27 subjects with a CDRS score of 0.5, 38 (91%) of the 42 subjects with a score of 1, and all 12 subjects with a CDRS score of 2.

<table>
<thead>
<tr>
<th>Individual Task Performance Scores Stratified by CDRS Score for All Subjects Who Completed the Testing Call*</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDRS Score</td>
</tr>
<tr>
<td>MMSE score</td>
</tr>
<tr>
<td>Orientation</td>
</tr>
<tr>
<td>Spell “FUN”</td>
</tr>
<tr>
<td>Immediate recall</td>
</tr>
<tr>
<td>Delayed recall</td>
</tr>
<tr>
<td>Directed key press</td>
</tr>
<tr>
<td>Auditory spatial relations</td>
</tr>
<tr>
<td>Backward digit span</td>
</tr>
<tr>
<td>Semantic comprehension</td>
</tr>
<tr>
<td>Sum of all IVR tasks</td>
</tr>
<tr>
<td>Call duration, mins</td>
</tr>
</tbody>
</table>

*All data are mean (SD) unless otherwise indicated. CDRS indicates Clinical Dementia Rating Scale; MMSE, Mini-Mental State Examination; and IVR, interactive voice response. Within each task, significant mean performance differences between groups (Newman-Keuls contrasts; \( \alpha = .05 \)) are indicated by the use of different superscript letters.

†F score for an analysis of variance of performances for each task.
SUBJECT PERCEPTIONS AND FEEDBACK

After completing the telephone call, subjects were asked for feedback about the IVR calling experience. They were asked to rate the overall difficulty of the telephone program and the difficulty of each of the tasks on a 5-point scale from “very easy” to “very difficult” with 3 indicating “neither easy nor difficult.” Not all of the subjects provided complete ratings. A total of 474 “task difficulty” ratings were provided by subjects with a CDRS score of 0, with 85% of these ratings indicating the system was easy or very easy to use, 8% indicating the system was neither easy nor difficult, and 7% indicating the system was difficult or very difficult. Subjects with CDRS scores of 0.5 or greater provided 436 such ratings, with 76% of the ratings being easy or very easy, 14% indicating the system was neither easy nor difficult, and 10% indicating the system was difficult to very difficult to use. Almost half (49%) of the difficult or very difficult ratings were given to the backward digit span task and another 29% given to the delayed recall task. In general, the subjects with CDRS scores of 0.5 or greater rated the tasks as more difficult than those with CDRS scores of 0, but the mean rating for all of the tasks for both groups was in the direction of easy to very easy. A total of 114 subjects answered a question about the clarity of task instructions provided by the IVR, with 93.9% indicating that the instructions were clear and allowed them to understand what they were supposed to do during the task.

IVR-ADMINISTERED INFORMANT SCREENING

Seventy-four collateral informants called the IVR system, entered an ID linked to a target subject, and responded to the 11-item SDS. The mean (SD) call length was 4 minutes 46 seconds (41 seconds). Of the 22 subjects with a CDRS score of 0, 13 had an SDS score of 0, 4 had a score of 1, and 5 had a score of 2. Of the 52 subjects with questionable, mild, or moderate dementia, 47 had an SDS score of 3 or greater. Standard scoring of the SDS (≥5) produced a sensitivity of 63.5% and specificity of 100%. These data suggest that using an SDS score of 3 or greater as a criterion might increase sensitivity to 90.4% without reducing specificity.

Objective computer-automated cognitive screening using IVR technology can discriminate between patients with early dementia symptoms and those without. The derived scoring model produced sensitivity and specificity estimates of roughly 80%. Application of the scoring model to data obtained from the validation sample supports generalizable validity. Adequate discrimination between cognitively impaired and unimpaired subjects did not require complete task performance, and the most discriminative tasks were those judged as easiest to complete by the subjects. This may partly reflect a sampling bias of impaired subjects with a mean CDRS score very close to 1. These data also indicate that collateral informants can use IVR technology to identify patients with early dementia symptoms.

COMMENT

To maximize benefits of current treatments for dementia, particularly for treatments of AD (which may only slow the symptomatic cognitive decline), early detection and recognition is critical. Wide-scale screening, whether through direct patient evaluation or collateral informants, poses significant challenges for support of the nec-
essary resources and logistics. This study demonstrates that IVR technology could play an important role in reliably identifying elderly patients beginning to manifest cognitive impairment suggestive of early dementia. Patients aged well into their 80s or 90s and even those with mild to moderate dementia can comprehend and navigate Touch-Tone interfaces to complete computer-automated assessments. The 10.3% hang-up rate is higher than desired; however, this problem can be reduced. Only 5 (3.2%) of the 155 subjects were unable to complete enough of the call to permit the application of a decision criterion that would have accurately identified their cognitive status. A total call length of about 12.5 minutes is not an excessive burden for accurate dementia screening; removal of unnecessary tests and use of a real-time scoring algorithm, terminating when a criterion for accurate classification was met, would decrease call length, task demands, and loss to hang-ups. Of the 148 subjects completing the call through just the orientation and spell FUN tasks, 50 had already met the criteria for a cogently impaired classification; 90% of these subjects had CDRS scores indicating questionable, mild, or moderate dementia (only 10% represented false-positive screens). Such results need to be replicated, but an accurate classification of a caller within the first few minutes of a call offers the potential to provide immediate feedback, education, and referral to local or national treatment resources.

In November 1999, public interest in a toll-free IVR system to provide dementia education and resource referral information was examined during a monthlong pilot study. Nearly 200 anonymous calls were received from a predominantly rural Midwest county of about 100,000 persons. These callers accessed information about dementia prevalence, risk factors, current treatment options, and local resources for treatment and caregiver support. Roughly half of the callers were concerned for a parent or grandparent and another 25% of the callers were concerned about themselves. Dementia screening using IVR was not available at that time, pending results from the present study.

The results obtained in this study must be viewed critically, pending further research and replication. Many factors related to patient and hardware variability will influence the reliability and validity of this type of telephone-automated testing. Certainly, physical disabilities (eg, severe loss of hearing or vision or disabling arthritis) or other neurological conditions directly influence an individual’s ability to understand task requirements and respond appropriately to IVR applications. Such considerations limit universal application of this approach to prospective patient screening, but most cognitively intact senior citizens are familiar with and able to navigate the many IVR systems that are increasingly being used by banking and government institutions, medical clinics, airlines, and cinemas. The diversity of shapes, sizes, and features available for telephone configurations rivals that of any other standard household equipment. Such variability of hardware instrumentation influences data reliability. This consideration was, in part, the reason that the tasks developed for this research focused as much as possible on cognitive performance and deemphasized psychomotor speed or response times. While all of the automated tasks incorporated a “time out” interval to allow for nonresponsiveness, the intervals were at least 3 seconds or longer, allowing some confidence that failure to respond was more likely a result of mental confusion than slowed motor response. The extent to which most of the subjects found the task demands in this study to be relatively easy supports this speculation. Use of a standard office telephone by all subjects at each of the study sites controlled for much of this type of instrumentation error variance in the present study and should be used for any future clinical or research use. The degree to which the use of any Touch-Tone telephone across individuals and/or over time would influence the reliability and validity of the type of data obtained by this system remains to be investigated.

Live clinician telephone interviews demonstrate acceptable convergence and reliability with face-to-face clinician assessment of behavioral and psychological symptoms of dementia. Interactive voice response technology has already been used for screening and diagnostic purposes in other medical domains. Kobak and colleagues demonstrated the sensitivity and specificity of an IVR mental health screener for identifying depressive and anxiety disorders, obsessive-compulsive disorders, and eating and alcohol use disorders. The present study extends the utility of this technology to dementia screening. Clearly, this type of computer-automated, remote-access technology cannot directly obtain sufficient information to permit differential diagnosis of different types of dementia. Such differentiation requires a wider assessment of patient histories, personal risk factors, and evaluation of medical tests by skilled clinicians. However, the sooner such assessments are made following the onset of abnormal cognitive difficulties, the greater the likelihood of obtaining maximum benefit from the appropriate course of treatment. Interactive voice response technology has been effective for providing patients with medical education from the comfort of their homes through the convenience of the telephone, and Mahoney et al demonstrated that this technology can effectively link community-residing elders with professional health care providers.

In conclusion, the pieces for a computer-automated telephone system that are able to provide integrated dementia screening, education, and treatment referral and monitoring services presently exist. As more effective treatments for AD and other dementias develop, economically efficient methods for identifying those in need and connecting them to treatment providers will become increasingly important in providing socially responsible and cost-effective care to the elderly.

Accepted for publication March 13, 2001.

This study was supported by grant 1R43AG16538 (Dr Greist) and grants AG15071 and grants AG17177 (Dr Rizzo) from the National Institute on Aging, Bethesda, Md.

Editorial comments provided by Warner V. Slack, MD, to a previous draft of the manuscript were very helpful in improving manuscript content and presentation. The assistance of Deborah A. Kaplan, Shelly Bierbaum, Robin Mitchell, and Jory Wilke, RN, KD, during data collection is also acknowledged and appreciated.
REFERENCES


