INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

Université d'Ottawa · University of Ottawa

THE

DEVELOPMENT OF

"CLINICALLY SENSIBLE" TOOLS

TO SCREEN FOR COGNITIVE IMPAIRMENT

IN COMMUNITY- DWELLING ELDERLY PERSONS

(Bridging the gap between Research & Clinical Practice by Balancing Discriminant Ability vs. Practicality)

Thesis submitted to
the Faculty of Graduate and Postdoctoral Studies
In partial fulfilment of the requirements for the
M.Sc. Degree in Epidemiology
University of Ottawa

Frank J. Molnar B.Sc. MDCM FRCP(C)

Clinical Associate, Division of Geriatric Medicine - The Ottawa Hospital

Member of the SCOHS Institute on Health of the Elderly in Ottawa-Carleton

Member of the Dementia Network of Ottawa - Carleton

Affiliate researcher, Loeb Health Research Institute

Defended 18 /Dec / 2000

ABSTRACT

Background; Despite its' prevalence and clinical relevance, cognitive impairment typically remains undetected in 50% of cases.

Objective; To develop clinically sensible (quick, simple, acceptable), accurate and readily recalled screens for cognitive impairment based on easily reproducible analytic strategies.

Methods; The Canadian Study of Health and Aging (CSHA-1) served as the derivation data set. 3MS cognitive screening questions which were judged as most likely to be employed by busy clinicians and which were significantly associated (via χ^2 analysis) with cognitive impairment were selected as independent variables for multivariate analysis. The screening tests derived from logistic regression and recursive partitioning analyses which most closely approximated the sensitivity and specificity of the entire 3MS were externally validated.

Results; Two logistic regression based scales and two recursive partitioning algorithms demonstrated sensitivities and specificities approaching those of the complete 3MS (approximately 80% and 60% respectively). The sensitivity was superior to that of the MMSE.

Conclusion; Readily reproducible multivariate analysis based strategies can be developed which generate practical screening tests with psychometric properties approaching those of the 3MS. Given the existence of verification bias, these screens as well as screens with higher sensitivity and lower specificity must be validated prospectively before they can be clinically employed.

ACKNOWLEDGEMENTS

This thesis has been presented at the following fora and incorporates feedback from attending physicians and psychologists :

(1) The Canadian Association of Gerontology (CAG) national conference (Oct 1999).
(2) The University of Ottawa Geriatric Medicine Academic Rounds (Jan 2000).
(3) The Élisabeth Bruyère Health Centre Research Rounds (April 2000).

I would specifically like to thank the following for making this thesis possible:

Thesis supervisors and committee members; Drs. I. McDowell, G. Wells, and M. Hing for their continued guidance, understanding and encouragement throughout the past years. Their ability to steer me back to the correct course (on numerous occasions) was greatly appreciated.

Drs. Dalziel, Power, Byszewski, Mitchell and Azad for providing the training necessary to care for people with cognitive impairment and for helping me understand the clinical implications of this research.

Thesis examiners, Drs. Stiell and Birkett, for their very thorough reviews. Their comments contributed significantly to the final version of this thesis. Dr. Spasoff for his moderation of the oral defence.

R. Aylseworth and B. Kristjansson for assisting me in my battles with the CSHA data.

The physicians at the Élisabeth Bruyère Family Medicine Centre who responded to my survey.

Steven Wright for helping me smile while struggling through the seemingly never ending articles which form the backbone of this thesis with his observation; "to steal ideas from one person is plagiarism - to steal from many is research". A unique way of highlighting how much we owe to those who came before us and to acknowledge the many authors listed in the references.

Most importantly, my family - Maureen, Margarita, Julie, Oscar and Bart - for helping me relax, laugh and place all things in their proper perspective. Life goes on.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES ***Page***

APPENDICES **Page**

1. INTRODUCTION

Cognitive impairment is a common and important symptom complex which may be secondary to a variety of etiologies including dementia, delirium and/or depression. Detection of cognitive impairment is important because it can have significant negative effects on clinical care, it may be associated with increased morbidity and mortality, and because certain causes of cognitive impairment, such as delirium and depression, may respond to treatment. The prevalence of dementia is escalating as our population ages. Delirium and depression occur more commonly in people with an underlying dementia. Consequently, the prevalence of depression and delirium can also be expected to increase substantially. Finally, the detection of dementia will become increasingly important as effective treatments continue to emerge. Detection of all forms of cognitive impairment is important for the care of individual patients and will become an increasingly important factor supporting the efficient functioning of our health care system.

Clinicians follow five basic steps in the care of patients at risk for cognitive impairment: (1) *screening* for cognitive impairment, (2) *diagnosis* of the specific cause(s) of the cognitive impairment, (3) *treatment* of reversible loss, (4) *management* strategies to compensate for irreversible loss and (5) *future planning* when management strategies can no longer safely compensate. This approach is supported by Rockwood's list of clinical questions clinicians should ask in assessing a patient with memory complaints [1]. This thesis will focus on the initial step of screening for cognitive impairment.

Cognitive screening tests can be divided into three broad categories; (1) brief clinical examinations of mental status, (2) short batteries of neuropsychometric tests, and (3) questionnaires based on the observations by informants of behavioural changes over a given time period [2]. I will focus on the first category, the brief clinical examinations of mental status, as these are the most readily transferrable to busy clinical practice. The neuropsychometric batteries require a psychologist as well as extra time and consequently are not useful in primary care [3]. The questionnaires based on observations of behavioural change are promising but will be reserved for future study.

After screening, the second step of assessment involves determining whether patients with cognitive impairment are suffering from dementia, delirium and/or depression. These diagnoses are not mutually exclusive but often coexist in individual patients. Subsequent steps include narrowing down the diagnosis of dementia to specific types of dementia (i.e. Alzheimer's Disease, Vascular Dementia, Mixed Dementia, Lewy Body Disease, Parkinson's Dementia, Frontotemporal Dementia, Primary Progressive Aphasia etc.). This step often involves brain imaging techniques (CT or MRI) as well as detailed neuropsychological evaluations which may last several hours. Throughout this process clinicians search for reversible components of the cognitive impairment, including depression and various causes of delirium. Once a diagnosis has been established then the clinician can move on to treatment (starting or stopping medications), management (community care, education and support for families, wandering registries, day programs for the patient...) and future planning (establishing power of attorney, encouraging patient to complete a will, determining resuscitation plans and advanced directives, etc.).

Screening tools exist for dementia, delirium and depression. Unfortunately these are of limited use unless one first suspects cognitive impairment based on history, the patient's behaviour or the results of a cognitive impairment screening tool. Detection of the cognitive impairment is the critical primary step which in turn triggers a more detailed diagnostic work-up. Research that focuses on screening tools for dementia or delirium misses this important first step. Such research assumes that clinicians can accurately detect cognitive impairment. This assumption is not supported by published literature which consistently indicates that physicians miss a significant proportion of cases of cognitive impairment.

There are further arguments supporting the need for the primary screen for the more general (i.e. non-specific) symptoms of cognitive impairment. Any form of cognitive impairment (i.e. not only dementia) can influence the ability to comply with treatment, the ability of patients to return home from inpatient settings, as well as the strategy to prevent future recurrences or exacerbations of a patient's medical condition. Some forms of cognitive impairment (e.g. delirium or depression) are more reversible than others (e.g. dementia). These more reversible forms of cognitive impairment (e.g. delirium and depression) might be missed if one only employed a screen for dementia. A similar argument holds for screens that focus only on depression or delirium - they will miss patients with the remaining forms of cognitive impairment. The separation of dementia, delirium and depression prior to the detection of cognitive impairment is an artificial approach which does not reflect the reality of clinical practice.

Unfortunately, even well known screens such as the Folstein Mini-Mental State Examination (MMSE) are not routinely employed in primary care due to various barriers such as time constraints or fear of upsetting patients [4]. There are no screens for cognitive impairment which physicians routinely employ in clinical practice. In the absence of such clinically practical screens, physicians have been shown to miss cognitive impairment in over 50% of cases [5-10]. The development of practical primary care screens for cognitive impairment will be the focus of this study.

The need for simpler and quicker screens for cognitive impairment with a higher degree of acceptability to patients and physicians has been previously identified. Past analyses have determined that no single cognitive question is adequate as a screen for cognitive impairment [11]. Multivariate analyses have produced predictive equations which unfortunately are not useful in clinical situations. A large gap continues to exist between clinical need and available screens for cognitive impairment. The gap has not been significantly narrowed by previous multivariate analyses because their results (i.e. the multivariate equations) lack clinical practicality.

A promising approach to narrowing this gap may be borrowed from Feinstein's concept of "clinical sensibility"[12]. Clinical sensibility describes whether a decision rule is clinically reasonable, easy to use (quick and simple to apply and interpret), and suggests a course of action.

Stiell and Wells have employed the concept of clinical sensibility in the development of their clinical decision rules for ankle fractures [13, 14]. This likely accounts for the widespread use of

these rules. Screening tools may be considered clinical decision rules which determine whether patients should go on to more in-depth evaluation. Consequently, it is reasonable to apply the concept of clinical sensibility to the derivation of screening tools. Application of the principle of clinical sensibility to decisions such as variable selection, variable coding, selection of the final format of the screening tool should enhance the ability of multivariate analysis to generate practical screening tools to detect cognitive impairment. This is the approach which this thesis will explore.

2. LITERATURE REVIEW

(2.1) The prevalence of cognitive impairment and dementia.

Dementing illnesses and cognitive impairment are highly prevalent in Canada. The largest population study of dementia and cognitive impairment to date was the Canadian Study of Health and Aging (CSHA)[15] . The CSHA was a multi-centre study of the epidemiology of cognitive impairment that involved a representative sample of 10 263 Canadians aged 65 and older from 36 cities and surrounding areas. According to the Canadian Study of Health and Aging, 8% of those over 65 and 34.5% of those over 85 demonstrate evidence of a dementia. Cognitive impairment is even more common since it may be secondary to dementia as well as other etiologies such as delirium, or depression. The CSHA estimates that 20.0% of community dwelling elderly are suffering from some form of cognitive impairment [16].

(2.2) The need for screening of seniors in primary care practice

In a survey of practicing family physicians, 95% of the physicians surveyed felt that cognitive impairment affects clinical management and 82% felt that cognitive screening is needed[4]. The 1991 Periodic Health Examination [17] found that there was insufficient evidence to include or exclude screening for cognitive impairment from the periodic health examination of people over 65 years of age (i.e. a C level recommendation). This is a situation in which we have a "lack of evidence of efficacy" rather than "evidence of lack of efficacy". This does not argue against the development of screens for cognitive impairment. In fact, it may not be possible to determine if cognitive screens are truly useful in primary care until a clinically practical screen is actually available for empiric study. In the absence of evidence for or against such screens, we must rely on clinical opinion. Consequently, the 1991 Periodic Health Examination guidelines list as a research priority the need to examine "the impact of screening for cognitive impairment and its subsequent investigation and treatment on the clinical outcome." The guidelines clearly support further research on the topic of screening for cognitive impairment.

Furthermore, the 1991 guidelines were written prior to the availability of effective treatments for dementia. Aricept (a.k.a. Donepezil) is an acetylcholinesterase inhibitor which has demonstrated statistically significant improvements in cognition in persons with mild to moderate Alzheimer's disease [18-20]. When the results of these trials are combined with the MCID (Minimally Clinically Important Difference) of Dementia Therapy study of Burback, Man-Son-Hing and Molnar (the candidate) [21], the data indicate that a significant proportion of patients are likely demonstrating not only a statistically significant but also a clinically significant

improvement in cognition. Since June 1999 Aricept has been covered by the Ontario Drug Benefits program and consequently is increasingly utilized in clinical practice. A second medication, Exelon, has recently been approved for the treatment of mild to moderate Alzheimer's disease. Given the fact that we now have two treatments for the most common form of dementia and given the fact that the evidence is based on double-blind randomized controlled trials (i.e. level 1 evidence), the level C recommendation of the 1991 Periodic Health Examination regarding screening for cognitive impairment is now outdated. It can be anticipated that the next update of the Periodic Health Examination will provide a level A or at least a level B recommendation to include a screen for cognitive impairment for persons over 65 years old. This underscores the pressing need for clinically practical screens for cognitive impairment.

(2.3) Present ability to detect cognitive impairment in primary care.

Past studies have demonstrated that physicians do not fare well when their ability to identify patients with cognitive impairment is tested [5-9]. In one study fewer than 50% of patients who were demented or delirious were recognized as such by their attending physicians. Two recent surveys suggest that this low detection rate may be due to the fact that physicians rarely formally test the mental status of their elderly patients [4,10]. It can be anticipated that this performance will continue to deteriorate as physicians are asked to see growing volumes of increasingly complex patients. The only feasible solution is to provide these clinicians with better education regarding the detection of cognitive impairment and to develop more practical screening tools.

(2.4) The routine use of available screening tools.

A number of tools to identify impaired cognition exist. These have been reviewed by Burns, Lawler and Craig [22] as well as by McDowell and Newell (refer to Appendix A on page 110) [23]. In their review of neuropsychological tests, Burns et al. pointed out that the majority of the neuropsychological tests are too long for use in the primary care setting. Burns points to three tests that might be useful in primary care; the Folstein MMSE, the Abbreviated Mental Test Score (ABMTS), and the Clock drawing. The briefest tests on McDowell's list are the Clock Drawing, Kahn's Mental Status Questionnaire (MSQ), Pfeiffer's Short Portable Mental Status Questionnaire (SPMSQ), and the Clifton Assessment Procedures for the Elderly (CAPE). Unfortunately the ABMTS, MSQ, SPMSQ, and CAPE each contain 10 - 12 questions. These screens are too long for busy clinicians to readily recall and routinely apply in clinical practice. This likely explains why these screens have never become popular in primary care. The Folstein MMSE and the Clock Drawing merit more detailed discussion.

The Clock Drawing is the simplest test to administer but is difficult to score. The patient is asked to draw the face of a clock depicting the time "ten minutes past eleven". This requires less than one minute to perform. Unfortunately the formal scoring is too complex to use in routine clinical practice (see Appendices B and C on pages 111 - 112). It is very rare for primary care physicians to routinely employ the Clock Drawing along with a validated scoring system. This is supported by Dr. Bush the chief neuropsychologist of the Élisabeth Bruyère Health Center Memory Disorder Clinic and author of a recent study of cognitive screening in primary care [4]: "Most neuropsychologists do not use a formal scoring method either [for the Clock Drawing].

I think it is totally unlikely that any physician would, as scoring any kind of drawing is the slowest kind of scoring there is..." (Personal communication). Informal scoring tends to be a gestalt sense (the "eyeball method" in Dr. Bush's words) that the drawing is poor enough to warrant more in-depth cognitive evaluation. Many physicians seem unaware that there are formal scoring methods for this test. This gestalt approach has never been empirically studied or validated but would likely demonstrate poor inter-observer reliability and would probably be insensitive to mild degrees of cognitive impairment. Regrettably, the complexity of the scoring systems have prevented the Clock Drawing from becoming a useful tool for screening in the primary care setting. The promise of the Clock Drawing as a rapid primary care screening tool is an illusion because it does not take into account the relatively complex scoring.

An innovative approach to the detection of cognitive impairment is the Time and Change Test [24] which incorporates a component of the clock drawing. Patients are asked to identify the time of a clock drawing set at 11:10, and then to select exactly one dollar in change from three quarters, seven dimes and seven nickels. The test has a sensitivity of 63% and a specificity of 96% for the detection of dementia in the outpatient setting. The authors reported that this is a "simple, accurate and reliable test". While the test does appear to be simple and reliable, it is not clear that it is accurate enough to serve as a screening tool. A sensitivity of 63% indicates that the test misses 37% of cases of dementia. This would likely be judged as having inadequate sensitivity by most clinicians. The selection of optimal sensitivity and specificity is a complex issue which will be discussed in greater detail. The structure of the Time and Change test score does not permit adjustment of cut-off scores in order to make the test more sensitive. Put

differently, it has a relatively rigid scoring system. Furthermore, the need to carry around this exact change (without spending it) is an anticipated barrier to widespread clinician use.

Perhaps the most widely used mental status test is the Folstein Mini-Mental State Examination (MMSE)[25] depicted in Appendix D on page 113. Numerous studies of the sensitivity and specificity of the MMSE have been performed. These have been reviewed by Tombaugh [26] and McDowell [23]. The MMSE is well established in the literature due to its numerous reliability and validity studies. A 1996 survey performed by the Geriatric Assessment and Intervention Network (GAIN) demonstrated that 95% of Canadian specialty geriatric services used the MMSE as their screen for cognitive impairment. Results of this survey are available at www.canger.org.

Unfortunately, even relatively simple and well known cognitive tests such as the Folstein MMSE are not widely employed by most clinicians in primary care. In Dr. Bush's study [4] only 24% of surveyed family physicians reported routinely screening their elderly patients for cognitive impairment. Almost all of these employed the MMSE. The authors felt that "social desirability bias" (i.e. giving the answer that is considered most correct) was a limitation of this survey and that the 24% likely overestimated the actual use of formal cognitive tests in clinical practice. While the MMSE is routinely used by specialty geriatric services (e.g. geriatric clinics, day hospitals, geriatric assessment wards, and outreach teams), it has not been adopted into routine use in the primary care setting. This is likely not due to lack of familiarity with the MMSE but rather due to other barriers which will be discussed in greater detail below.

In summary, there are no mental status tests or cognitive screens which have been widely adopted into routine primary care practice. In order to understand why this is true we must review the barriers present in the primary care setting. Only by directly addressing such barriers can we develop a screen which might become more widely employed.

(2.5) Barriers to screening in primary care settings.

The survey by Bush et al. indicated that the large gap between perceived need and actual practice is largely due to the time constraints of clinical practice which preclude spending the 5 - 15 minutes required to administer the MMSE. If this short period of time is multiplied by the number of patients over 65 seen daily, the time requirements rapidly become prohibitive for most family physicians. Fear of upsetting patients with a long cognitive test was cited as another barrier to the routine use of cognitive screens. Bush, Kozak, and Elmslie concluded that a "simple, quick screening test that is acceptable to both physicians and patients" is required [4]. This appears to incorporate the concept of clinical sensibility promoted by Feinstein [12], and by Stiell and Wells [27].

(2.6) Past Bivariate Analyses to determine the best cognitive screening question.

In response to this need for a more rapid screen, individual MMSE items have been analyzed [11] . Such studies have demonstrated that there is no single question with adequate sensitivity and specificity to permit it to stand alone as a screening test .

Siu [11] has concluded that "responses to a *combination* of these brief screening questions can best be used to determine the need for additional mental status examination". Unfortunately, Siu did not elaborate on this approach. To date, only very preliminary studies employing logistic regression have been performed to select the optimal subset of questions. The results have been somewhat disappointing as will be discussed below. It is the goal of this thesis to further this specific line of research.

(2.7) Past Multivariate Analyses to determine the best set of questions.

Several investigators have examined shortened versions of the MMSE [28-32]. Paveza [28] and Koenig [29] studied the first half of the MMSE questions which represent the verbal portion. This represents a "convenience sample" of questions (i.e. the easiest to complete) rather than an analysis to determine the best possible subset of questions. Other studies [30-32] have employed multivariate statistical analysis to identify the best subset of cognitive questions.

Klein [30] employed multivariate analysis of MMSE items to determine which were the most predictive for dementia. Unfortunately, the sensitivity and specificity reported by Klein represent those of the variables when they are incorporated into the complex mathematical equation generated by multivariate analysis. It is unrealistic to expect physicians to be able to incorporate such complex equations into clinical practice. In addition, Klein focused on dementia rather than cognitive impairment. As discussed previously, it is also important to identify non-dementia related cognitive impairment such as depression or delirium since these are potentially even more reversible than dementia and are associated with high levels of morbidity & mortality.

Magaziner [31] performed a similar regression analysis; however, these results were even less practical from a clinical perspective. The study generated four different regression equations. The particular equation selected would depend on the patient's age and educational level. It is exceedingly unlikely that clinicians would be willing to recall and employ four different equations in routine clinical practice. In addition, Magaziner's equations included a test that requires pencil and paper (an anticipated barrier to use by physicians) as well as a total of seven questions which would likely be judged as too lengthy for most clinicians to remember and use.

In 1992 Braekhus [32] employed logistic regression analysis to determine which MMSE items were most predictive of cognitive impairment as determined by the overall MMSE score. A simple scoring system was developed which could be used at the bedside. While his analysis was quite promising due to the simplified scoring, the results once again lacked clinical practicality. The subset generated included a question that requires pencil and paper. Even more of a barrier to widespread use, the final subset included 12 questions which is much too long for clinicians to recall.

Clearly, a large gap persists between the screening tools being generated by researchers and the needs of clinicians. Research to date appears to have stalled at this point without having produced the brief practical screens required by clinicians. Fortunately, there exist other promising approaches which this thesis will explore.

(2.8) New considerations in the development of practical screening tools for cognitive impairment

[2.8.1] Clinical "Sensibility"

At each step in the generation of cognitive screens we must balance test characteristics (sensitivity, specificity, overall accuracy) with the needs of clinical practice. There is a natural tension between the competing desire to maximize the discriminant ability of a test (which often means making the test longer and more complex) and the need to make the test practical in a clinical setting (which means making the test as brief and simple as possible). These competing demands are reflected in the concept of "clinical sensibility". Feinstein [12] first used the term "sensibility" to describe whether a decision rule is clinically reasonable, easy to use and suggests a course of action. Stiell and Wells [27] have pointed out that "ease of use depends on factors such as the length of time needed to apply the rule and simplicity of interpretation". They also note that "assessment of sensibility depends on judgement rather than on statistical methods". Incorporating the principle of clinical sensibility into variable selection, variable coding, tool format, tool length and scoring decisions should maximize the probability that a screen for cognitive impairment will be routinely employed in clinical practice. We will review methods of incorporating sensibility into the design of practical screening tools for primary care.

[2.8.11] Selection of variables.

To date, multivariate analyses have been performed on all variables available. There has been no pre-selection of variables. If one wishes to develop a screening test that may be widely used, the logical modification of the approach is to pre-select questions that busy clinicians are

most likely to find acceptable. Perceived barriers to the incorporation of cognitive screening questions into clinical practice include the need for paper, pencil, cue cards or other props. Clinical experience indicates that physicians are much less likely to routinely use a cognitive test if these are required. A previous (1997) review of open charts of patients being followed by the Ottawa Civic Hospital Geriatric Medicine service indicated that 12% of day hospital patients (outpatients) and 42% of ward inpatients did not complete the full MMSE. The initial verbal component was completed in almost all instances. The sections that were often not completed include; [1] phrase repetition, [2] read and obey "close your eyes", [3] write a sentence, [4] copy intersecting pentagons and [5] three stage command.

Physicians appear to be most comfortable with questions requiring verbal (as opposed to written) responses possibly because the former parallel standard medical history taking techniques. However, focusing on questions requiring verbal responses may limit testing of certain cognitive domains such as visual-spatial function. This will be discussed in greater detail.

[2.8.12] Coding of variables

Another barrier to the incorporation of a screen into routine practice is complexity of scoring. The clock drawing is a prime example. Here we are faced with competing disciplinary approaches. Neuropsychologists utilize more complex scoring systems which permit more detailed gradations of "correct". They also evaluate complex qualitative features of the patient's performance such as difficulty in answering a question, speed of response etc., and they place this within the context of the patient's expected baseline performance given their level of

education and occupation. Such complex evaluations are not feasible during the rush of medical practice. Consequently, front-line physicians tend to dichotomize results into right/wrong or pass/fail rather than into multiple categories [33]. Past studies have not taken this into account but have maintained the original scoring system for each cognitive question. This can include items such as a 10 point scoring system for the intersecting pentagons of the Modified Mini-Mental Examination (refer to Appendix E on page 114). It is exceedingly unlikely that busy clinicians will be able to recall and employ scoring systems of this level of complexity.

In order to maximize the probability that physicians will routinely employ a screening tool in clinical practice one should incorporate a scoring system that the clinicians will find most natural. In our case this will involve dichotomization of the results of each cognitive question; the patients are either correct or incorrect in their responses. It is recognized that discriminant power will be lost when one moves from continuous to categorical and then to dichotomous scoring.

[2.8.13] Format and scoring of the screening tool.

Simpler scoring systems than that provided by a regression equation are also required in clinical settings. Complex multivariate equations are not useful at the bedside. As Stiell and Wells have indicated; "In the emergency department, it is unlikely that physicians would embrace a rule that requires extensive calculations or the use of a calculator"[27]. This is equally true of physicians in all other settings. There are no existing examples of the successful application of a multivariate equation to routine clinical practice in the primary care setting.

A useful approach may be borrowed from the scoring system employed by the CAGE questionnaire [34,35] (see Appendix F on page115). When tests (in this case cognitive questions) are combined, physicians are readily able to use simple "unweighted" scales where each question contributes equally to the final score. This is demonstrated by the widespread use of the CAGE questionnaire which has the added benefit of a simple acronym to help in the recall of the questions. In the area of selection of questions to screen for cognitive impairment, the first 4 cognitive questions (independent variables) selected via logistic regression could each be assigned a value of 1 point. The sensitivities and specificities of various cutoff scores (<4, <3,<2) could be determined and receiver operating characteristic (ROC) curves could be generated. These steps, while methodologically straightforward and sensible, have not yet been pursued in the published literature on cognitive impairment.

The "CAGE approach" generates an unweighted scale where each of the selected variables contributes equally to the overall score. While this would result in a simple and clinically usable scale, a great deal of information (i.e. discriminant ability) may be lost. A compromise might be to use the coefficients generated by the logistic regression equation to weight variables in the scale. This was the approach employed in the generation of the Goldman Cardiac Risk index [36] (see Appendix G on page 115). Unfortunately, this technique may result in a scale which is already too complex to be recalled and thus will not be widely used at the bedside. Many clinicians use the items of Goldman's Cardiac Risk index but very few actually apply the weights in day-to-day practice. The weighting is a barrier to widespread use as physicians must carry a copy of the scale with them. Even less practical is Detsky's modification

of the Goldman index (see Appendix H on page 116) which requires that physicians carry a nomogram with them at all times [37]. It has also been suggested that the coefficients derived from the multivariate equation may be very sensitive to changes in population (i.e. minor changes in spectrum of disease which will be discussed later) and may not be stable [38]. Overall, the concept of a weighted scale deserves more study to determine if the weighting adds significantly to the sensitivity and specificity of the screen and if these gains outweigh the possible loss of acceptability to clinicians.

In summary, the more complex the scoring system employed, the less likely it is that a particular screen will be used clinically. Conversely, the more one simplifies a screening test, the more discriminant ability one loses. Clearly one must balance competing needs (i.e. practicality vs. discriminant ability). This tradeoff is one which this study will explore.

[2.8.14] The optimal length of a screening tool in primary care.

Previous studies have indicated that no single question is adequate as a screen for cognitive impairment [11]. Recent research suggests that while physicians appear to know the four CAGE questions, they seldom ask patients all four [35] . This indicates that four questions may be at the upper limit of what clinicians will adopt into routine practice. As one decreases the number of questions to 3 or even 2 the test may gain acceptability but once again this comes with the price of a loss of discriminant ability. The above suggests that 2 to 4 questions may be the range of the maximum acceptable number of questions for a quick and simple primary care screening tool. More research into the optimal length of clinical screening tools is justified.

As the number of screening questions is increased, the ability to recall the questions and hence the acceptability of the tool is rapidly lost. In multivariate analysis each new independent variable added to the equation accounts for less variability than previous variables (i.e. cognitive questions in this case). Given this "law of diminishing returns" and the loss of acceptability as a tool is lengthened, it becomes increasingly difficult to justify lengthening a screening tool.

[2.8.2] Selection of appropriate analytic techniques for the construction of cognitive screens.

The dependent variable in this study is dichotomous; "cognitively impaired" vs. " not cognitively impaired". Consequently, the relevant analytic techniques include (1) discriminant function analysis, (2) logistic regression, (3) recursive partitioning and (4) cluster analysis. The latter is beyond the scope of this thesis as it is more suited to distinguishing among several groups where there is no dependent variable. The first three analytic techniques are more appropriate for the development of cognitive screens and will be discussed in greater detail.

[2.8.21] Discriminant Analysis vs. Logistic Regression.

The derivation of the Goldman cardiac risk index [36] was based on multivariate stepwise linear discriminant analysis. In the cognitive field, Storandt [38] and Eslinger [39] have employed discriminant analysis to develop a weighting system to improve the discriminant ability of neuropsychological tests. Unfortunately the weights proved to be very sensitive to changes in population and, consequently, were unstable.

While discriminant analysis permits the weighting of variables, it unfortunately requires a normal distribution of the dependent variable. Hosmer and Lemeshow [40] have stated that "discriminant function estimators are sensitive to the assumption of normality.... the practical implication of this is that for dichotomous independent variables the discriminant function indicators will overestimate the magnitude of the association." Given the fact that in this thesis all independent variables have been dichotomized and the dependent variable is dichotomous, discriminant function analysis would not be appropriate. Furthermore, Braekhus et al. [32], and Ritchie and Fuhrer [2] have demonstrated that the raw data from mental status tests (i.e. the continuous data which has not yet been converted to dichotomous data) do not have normal distributions and consequently it would not be appropriate to employ discriminant analysis even if dichotomization were not employed to simplify scoring.

Logistic Regression analysis is ideally suited to our situation in which the dependent and independent variables are dichotomous. This is reflected by the selection of logistic regression in all recent similar studies attempting to find the best subset of questions to detect cognitive impairment.

There is, however, an inherent limitation of standard logistic regression analysis as it has been employed in the field of cognitive research. Logistic regression algorithms do not maximize either sensitivity or specificity but instead attempt to maximize the "total % correct" which is a combination of sensitivity and specificity. Put differently, the software is designed to maximize accuracy which may not always be the optimal approach in the generation of

screening tests [27]. It may be possible to vary the prevalence of disease or to create a weighted analysis which would either maximize sensitivity or specificity. This approach has not been pursued in the published literature on the development of cognitive screens but will be explored.

[2.8.22] Recursive Partitioning

Maximizing sensitivity is consistent with the recursive partitioning approach employed in the derivation of the Ottawa Ankle Rule studies [13, 14] and the Ottawa Knee Rules (see Appendix I on page 117). Recursive Partitioning [41] appears to be well suited to the goal of maximizing either sensitivity or specificity. Chi-square (χ^2) recursive partitioning techniques may be used to create a regression tree with splits based on dichotomous responses to cognitive questions. This form of analysis has also never been explored in the field of cognitive research.

3. OBJECTIVES

Primary Objective

To develop clinically sensible (quick, simple, acceptable), accurate and readily recalled screens for cognitive impairment in unselected community dwelling seniors. These are to be based on the simple and easily reproducible analytic strategies discussed below.

Secondary Objective

To develop simple and easily reproducible analytic strategies, which balance accuracy and practicality, that can be used to generate clinically sensible screening tools for cognitive impairment.

4. METHODS

(4.1) Overview of the basic analysis (a more detailed description follows).

The Canadian Study of Health and Aging (CSHA 1 & 2) data bases were selected as they are representative of the target population (i.e. unselected community-dwelling seniors). These data bases are well matched and consequently it should be appropriate to employ the CSHA-1 data as the derivation set and the CSHA-2 as the main validation set.

The diagnosis (i.e. the dependent variable) was dichotomize into cognitively impaired or cognitively normal. Independent variables (i.e. cognitive screening questions) were selected based on their anticipated acceptability to primary care physicians (i.e. based on clinical sensibility). Consequently, questions that may create barriers to widespread use of the final screening tool such as questions requiring props (paper, pencil, cue cards), questions requiring greater than 30 seconds to answer, and questions that cannot be easily scored in a dichotomous fashion, were excluded from analysis. All responses to the remaining independent variables (i.e. cognitive questions) were dichotomized - the responses were coded as correct or incorrect.

Chi - square (χ^2) analysis of the selected dichotomously scored independent variables was performed (employing cognitively impaired vs. cognitively normal as the dependent variable). Next, multivariate analysis (logistic regression and recursive partitioning) of the independent variables found to be significant via bivariate (χ^2) analysis was performed.

Weighted logistic regression analysis was performed to derive more sensitive or, alternatively, more specific equations. The results of the logistic regression analyses were translated into a clinically usable format. This involved creating weighted (by Beta coefficients) and unweighted scales from the first 2,3 and 4 variables selected by logistic regression. The sensitivities and specificities of all possible cutoffs for the derived scales were calculated and ROC curve analysis was performed.

Similarly, recursive partitioning analysis was performed using the independent variables found to be significant via χ^2 analysis. Once again, the sensitivities and specificities of all possible cutoffs for the derived algorithms were calculated.

External validation of the scales and algorithms that have the best combination of sensitivity and specificity (in this case those that approached the properties of the 3MS) was performed using the CSHA-2 data as the primary validation set and the Ottawa General Hospital Memory Disorder Clinic data as the secondary validation set.

The above represents a very basic and simplified outline of the analysis. Selected analytic steps will be described in greater detail below.

[4.1.1] The databases

Due to the desirability of using compatible data sets for screen development and testing, the derivation set and primary validation set have been changed from those initially proposed.

[4.1.11] The Canadian Study of Health and Aging: derivation and validation data bases

The Canadian Study of Health and Aging (CSHA-1) [15] performed in 1991-1992 was a multi-centre, population-based cohort study of dementia with a sample of 10,263 participants aged 65 or over. The sampling procedure was relatively complex. Similar numbers of participants were included from five regions (e.g. Atlantic provinces, Quebec, Ontario, Prairies and British Columbia) which does not reflect population distributions. The older groups were over-sampled: the sampling fraction among subjects 75 - 84 was twice that among those 65 - 74, and the fraction among those 85 and older was 2.5 times that among those 65 - 74.

Of the community dwelling sample, the 1600 who screened positive for cognitive impairment (3MS score < 78) and a random sample of 494 who screened negative (3MS score > 77) received a full clinical evaluation. This full evaluation involved a detailed clinical and neuropsychological examination and finally a consensus diagnosis. This evaluation may be considered a relatively good criterion standard or gold standard which is far superior to that used in prior studies (i.e. often the MMSE score was the criterion standard in the older studies). The subset of the community sample who underwent full clinical assessment would be suitable as the population for the derivation of a screening test for cognitive impairment in the community.

The CSHA - 2 was a follow-up study performed in 1996-1997 of the surviving CSHA-1 patients. Since both the CSHA-1 and the CSHA-2 studies employed the same approach to screening and evaluation within the same population, the CSHA-2 data is suitable as the validation set for screening tests derived from the CSHA - 1 data.

[4.1.12] The Ottawa General Hospital Memory Disorder Clinic database.

Demographic and diagnostic data were collected via a chart review of 452 consecutive patients presenting to the Ottawa General Hospital Memory Disorder Clinic between 1988 and 1994. Patients were primarily referred by community based physicians. Assessments were completed by a team comprised of a geriatrician, a psychiatrist and a neuropsychologist. Evaluations included a thorough history, physical examination, blood-work, EEG, CT scan of the head, and detailed neuropsychological examination. Final diagnosis was based on consensus opinion of the entire team. This diagnostic approach is very similar to that of the CSHA.

The data collection strategy and forms were developed by the candidate (Dr. Molnar). Data were collected by a medical student under the supervision of the candidate and Dr. Hing. The Ottawa General Hospital Memory Disorder Clinic database was created by the candidate. Diagnoses were dichotomized into "cognitively impaired" vs. "not cognitively impaired". The data collector was blinded to the final diagnosis during collection of neuropsychological data.

The patients in this data set do not represent those of primary practice due to referral bias which has likely altered the natural spectrum of disease. While the CSHA employed 3MS data, this clinic only provided data from the MMSE. This data set is not well matched to the CSHA-1 data. Consequently, this database will not be used as either the derivation or the primary validation set, but will be used as a second validation set. If the screening tools developed in this study also perform well in this secondary data then this will provide further support for the screening tools as this would demonstrate increased external validity.

[4.1.2] Selection of subjects (i.e. inclusion and exclusion criteria)

Given the goal of developing a tool to screen for cognitive impairment in the community, institutionalized seniors were excluded. Other exclusion criteria included primary language other than English, sensory deficits (i.e. hearing, vision) severe enough to affect cognitive testing, and severe cognitive impairment (MMSE < 11). Finally, people were selected from the remaining subgroup if they had undergone full medical and neuropsychological evaluation since the diagnoses derived from the latter served as the gold standard.

[4.1.3] The dependent variable; selection of an appropriate criterion standard

Previous analyses [32] attempting to determine the best subset of cognitive screening questions have selected their subsets of questions from the Folstein Mini-Mental State Examination (MMSE) and have then used the full MMSE as the citerion to determine the sensitivity and specificity of the subset. This is an exceedingly weak design and is possibly inappropriate. Using the MMSE as the criterion for the subset of questions derived from the MMSE is effectively a study of the internal consistency of the MMSE items or, put differently, a measure of redundancy of information provided by the different questions. This part-to-whole comparison will give spuriously high correlations and does not address the underlying issue of which questions best identify people independently classified as cognitively impaired.

A superior criterion standard would be the consensus diagnosis reached by a team of clinical experts after a thorough review of the history, physical examination, blood-work, CT scan of the head and neuropsychological testing. Such high level evaluations were performed in

the CSHA-1, CSHA-2, and the Ottawa General Hospital Memory Disorder Clinic. This is likely the best criterion standard that can be expected and can be considered the gold standard.

Consequently, the dependent variable was the consensus diagnosis of "cognitive impairment" vs. "no cognitive impairment" as determined by the CSHA staff or the Ottawa General Hospital Memory Disorder Clinic team (based on the full medical and neuropsychological evaluation).

[4.1.4] Selection and scoring of independent variables (i.e. cognitive questions)

Most published studies have selected cognitive questions from the well known 30 point Folstein Mini-mental State Examination. The Modified Mini-Mental State Examination (3MS examination)[42] is a 100 point cognitive screening test which has the majority of the smaller MMSE imbedded within it. The 3MS as displayed in Appendix E (Page 114) may be compared with the MMSE in Appendix D (P. 113) . The 3MS has been found to have superior psychometric properties when compared with the MMSE [43]. Selecting subsets of cognitive questions from the longer 3MS should yield a wider variety of possibilities than if the MMSE were used as the source of the questions. Consequently, cognitive questions were selected from the CSHA Modified Mini-Mental State [3MS] cognitive questions.

In order to maximize clinical acceptability, it was decided that selected cognitive questions should take no longer than 30 seconds each. This resulted in exclusion of the listing of four legged animals which can take up to one minute if one factors in time for explanation and

scoring. The population norm for this test is often not recalled by clinicians. Based on the previously discussed review of MMSE items completed in open charts at the Ottawa Civic Hospital, it was also decided that selected cognitive questions should not involve the use of pencil, paper, calculators or cue cards. This resulted in exclusion of; (1) phrase repetition, (2) read and obey "close your eyes", (3) write a sentence, (4) intersecting pentagon drawing, and (5) three stage command.

Independent variables were selected for analysis based on these a priori practical criteria. Responses to each question were dichotomized into right/wrong or pass/fail to promote ease of physician recall and scoring.

All independent variables (i.e. dichotomized 3MS cognitive questions) which met the practical selection criteria and which were found to be significant via bivariate analysis (chi - square) were entered into logistic regression (forward LR algorithm) and recursive partitioning analyses.

(4.2) Logistic Regression Analysis

SPSS software (version 9.0) was used for the logistic regression analysis (forward LR algorithm). Since the logistic regression software maximizes overall accuracy (rather than either sensitivity or specificity), an incremental weighting scheme was examined to see how this would alter the selection of variables in the equations. Cases (i.e. cognitively impaired) were

incrementally weighted from 1 to 10 with controls (i.e normals) maintained at a weight of 1. Logistic regression analysis was run for each weight ratio (e.g. cases-controls = 1-1, 2-1, 3-1,4-1, 5-1, 6-1, 7-1, 8-1, 9-1, 10-1) to create more sensitive equations. The higher the weighting on the cases, the more likely the equation will include the most sensitive independent variables (i.e. cognitive questions). Next, controls were incrementally weighted from 1 to 10 with cases maintained at a weight of 1. Logistic regression analysis was run for each weight ratio (e.g. cases-controls = 1-1, 1-2,1-3,1-4,1-5, 1-6, 1-7, 1-8, 1-9, 1-10) to create more specific equations. The higher the weighting on the controls (i.e. the cognitively normal), the more likely the equation will include the most specific cognitive questions.

The first four questions which were selected by the forward LR algorithm were then entered into logistic regression equations containing 2, 3 and 4 variables to determine how this would alter the beta coefficients.

Simple scales comprising 2,3 and 4 unweighted questions were created from the first 4 independent variables selected by Logistic Regression. These variables represent the 4 most predictive cognitive questions. "Unweighted" refers to the fact that each question was worth 1 point and contributed equally to the overall score. This is similar to the scoring of the CAGE questionnaire (Appendix F on P. 115). Next, more complex scales using the rounded-off logistic regression equation beta coefficients as weights for each question were created. This generated weighted scales which were similar to the Goldman Cardiac Risk Index (Appendix G on P. 115).

The sensitivity and specificity of every possible cut-off score for each of the above scales were calculated. For simplicity, where weighted and unweighted versions provided equivalent results, the unweighted scales were retained. Receiver operating characteristic (ROC) curve analyses were performed and the area under the ROC curve (AUC) was calculated.

(4.3) Recursive Partitioning Analysis

The same independent variables (i.e. cognitive questions) employed in the logistic regression analysis were selected for recursive partitioning. In order to develop cognitive screens via a reproducible methodology (see objective #2) it was decided to base the algorithms on the test properties which may be derived from a 2 X 2 table (refer to Appendix J on Page 118).

The χ^2 recursive partitioning software "Knowledge Seeker" was employed. Algorithms which selected questions based on the following criteria were developed: (1) maximally sensitive questions [this parallels what was done in the derivation of the Ottawa Ankle Rules]; (2) maximally specific questions; (3) questions with the maximal Positive Predictive Value (PPVmax); (4) questions with the maximal Negative Predictive Value (NPVmax); and (5) questions with the maximal accuracy. Given the secondary objective of generating simple and easily reproducible analytic strategies, the same criterion was used to select each question in a particular algorithm (e.g. one algorithm would use only the maximally sensitive questions, another would use only questions which had the maximal PPV etc.). Since the spectrum of disease is altered after each split in a recursive partitioning tree; the sensitivity, specificity, PPV, NPV and accuracy must be recalculated for all remaining questions after each split.

The above procedure has the merit of simplicity. It is recognized that algorithms that attempt to use combinations of these properties (e.g. maximal sensitivity for the first question, PPV for the second question, NPV for the third ...) might have superior psychometric properties but would not result in a simple and easily reproducible methodology. The process of deriving such hybrid algorithms would likely be very researcher dependent and would not be readily reproducible. Consequently, the hybrid approach was not attempted in this study.

In each of the above five recursive partitioning analytic schemes (Sensitivity max, Specificity max, PPV max, NPV max, and Accuracy max) the second, third and fourth questions can always be applied only to people who passed the previous question. Alternatively, the questions can be applied only to people who failed the previous question. Refer to figures 1 and 2 (Page 102) for examples of these alternate directions / approaches. Consequently, there are 10 basic recursive partitioning schemes which can be derived from a standard 2x2 table (i.e. [5 criteria - sensitivity, specificity, PPV, NPV and accuracy] X [2 algorithmic directions depicted in figures 1 and 2]). This thesis was limited to the exploration of these ten basic algorithms.

(4.4) Selection and notation of cut-off scores

In the absence of guidelines regarding the optimal sensitivity and specificity of a screen for cognitive impairment, cut-off scores were selected to match the psychometric properties of the published cut-offs of the 3MS as closely as possible. The goal was to find a cut-off score with a sensitivity equivalent to that of the 3MS and then to maximize specificity. The complex issue of selection of cut-off scores will be reviewed in more detail in the discussion section.

Throughout this study the terms cut-point, cut-off score and threshold will be used interchangeably. While there is no universally accepted notation to describe cut-off scores the following are the two most commonly employed ; A/B or A-B. This is interpreted as "any number less than or equal to A is abnormal (i.e. cognitively impaired) and any number greater than or equal to B is normal (i.e. cognitively intact)".

(4.5) ROC Analysis

The ROC Analyzer (software copyrighted by R. Centor and J. Keightley 1992) was employed. This software utilizes the maximum likelihood estimation (MLE) technique to determine the area under the curve (AUC). The MLE fit assumes binormality (i.e. that both the normals and the abnormals have ratings which derive from underlying normal distributions). Published literature has repeatedly demonstrated that this assumption is false in the case of cognitive impairment. Ritchie and Fuhrer [2] found the cognitively normal group had a skewed distribution of cognitive scores and the demented group in their study had a platykurtic distribution of cognitive scores. Consequently such parametric analysis is not appropriate for cognitive screening data. In this study all the data have been dichotomized. This automatically rules out the parametric option. Conveniently, the ROC Analyzer software also performs the nonparametric analysis of Hanley and McNeil (i.e. the trapezoidal method). This form of ROC analysis was selected. As the results will demonstrate, the nonparametric analysis will provide a more conservative (i.e. smaller) estimate of the AUC and may even underestimate the true AUC. Selection of the Hanley and McNeil trapezoidal method is supported by previous research in the field of cognitive screening [2, 43].

(4.6) Validation of derived cognitive screening tools

The above is essentially an exploratory data analysis and both methods will tend to over-fit the models to the study samples. Consequently, validation is required. Split sampling with cross validation, jack-knife and bootstrap analyses are possible but validation in another data set (i.e. external validation) is a superior option. In this case the CSHA-2 data was chosen to serve as the primary external validation set and the Ottawa General Hospital Memory Disorder Clinic data served as a secondary validation set.

Prediction rules often do not perform as well in a new set of patients or in the real world setting. External validation (in another data set) tests the former but not the latter. Prospective validation is superior to all of the above forms of validation as it tests effectiveness in actual clinical practice. Unfortunately, prospective validation lies outside the scope of this thesis.

(4.7) Pilot Survey of the acceptability of different screens to practicing physicians

Family physicians practicing at the Élisabeth Bruyère Family Medicine Centre were surveyed to determine how acceptable they considered the cognitive screens developed in this study in comparison with the MMSE and the clock drawing (the latter in conjunction with a formal scoring system). They were asked to estimate the likelihood that they would use each form of screening test in their everyday clinical practice. They were also asked to rank the screens in order of acceptability. The survey is displayed in Appendix K (Pages 119 - 125). This should be considered a pilot survey as developing tools to accurately assess acceptability of screening formats is a complex area which could serve as a thesis topic in its own right.

5. RESULTS

(5.1) The data bases

The data bases differed in terms of average age, percentage of cognitively normal patients, and education. The first two differences were expected given the origins of the three data bases. The average cognitive scores, however, were similar.

TABLE 1; Demographic and clinical comparisons of the data sets

	CSHA-1	CSHA-2	OGH Memory Disorder Clinic
Number of subjects	1134	1467*	452
mean age (SD)	79.5 (6.7)	81.9 (6.8)	70.6 (10.9)
% female	60.8	56.6	61.7
% cognitively normal	53	50	32
Education %			
- no formal schooling	.6	.4	.2
- primary school	39.8	33.5	15.3
- some high school	24.2	25.8	29.0
- finished high school	23.7	30.9	20.1
- post-secondary	10.0	6.9	21.7
- missing data	1.7	2.5	13.7
mean MMSE	23 - 24**	24 - 25**	23.1
3MS	76.8	77.7	74 - 76***

* larger number of subjects in CSHA-2 due to less selection within this data base
** the mean MMSE was calculated / estimated based on the mean 3MS value
*** the mean 3MS was calculated / estimated based on the mean MMSE

(5.2) Psychometric properties (sensitivity and specificity) of the MMSE and the 3MS in the detection of cognitive impairment in the CSHA-1 community dwelling population.

[5.2.1] The MMSE

McDowell, Kristjansson, Hill, Hébert [43] employed a cut-off of 25/26. In the CSHA-1 data this cut-off has a sensitivity of 70% and a specificity of 73% for mild to moderate cognitive impairment. The AUC for the MMSE in the detection of cognitive impairment was 0.77.

[5.2.2] The 3MS

Two cut-off scores have been published. In the absence of clear guidelines regarding the optimal sensitivity and specificity of screens for cognitive impairment, the goal of this study is to develop a screening tool and select cut-points which will approximate the performance of the 3MS. Teng and Chui [42] employed a cut-off of 79/80. In the CSHA-1 data selected for this thesis this cut-off has a sensitivity of 86% and a specificity of 60% for mild to moderate cognitive impairment. McDowell, Kristjansson, Hill and Hébert [43] employed a cutoff of 77/78. In the CSHA-1 data employed in this thesis, this cut-off has a sensitivity of 84% and a specificity of 62% for mild to moderate cognitive impairment. The AUC for the 3MS in the detection of cognitive impairment was 0.80.

(5.3) Bivariate analysis of individual cognitive questions

Based on the practical (i.e. clinically sensible) criteria outlined in section 4.1.4 on Page 27, twenty-four questions were selected as independent variables. Bivariate analysis (χ^2) was performed. Those twenty-one variables which were significantly correlated with the dependent variable were included in the multivariate analyses. Sensitivities and specificities were calculated to determine if any single question could serve as a cognitive screen on its own. The results of the bivariate analysis depicted in Table 2 (P. 88) demonstrate that no single cognitive question has a combination of sensitivity and specificity approaching that of the 3MS.

(5.4) Multivariate analysis (logistic regression)

When the forward LR algorithm was employed with no limitation on the number of variables entered along with a 1:1 weight ratio (case: controls), a 12 variable equation was generated. The equation demonstrated the following properties: Sensitivity = 61 %, Specificity = 81 %, and Accuracy = 71 %.

To date the literature on the selection of the optimal subset of cognitive screening questions has not moved beyond the generation of such lengthy and impractical logistic regression equations. All ensuing results and discussion represent original contributions to this field.

(5.5) Selection of the optimal subset of questions

[5.5.1] The Logistic Regression approach

It should be reiterated that the SPSS logistic regression software maximizes overall accuracy rather than sensitivity or specificity. By varying the weighting on either cases of cognitive impairment or on normals (i.e. controls) one may generate equations and hence select sets of questions which are either more sensitive or more specific. The incremental weighting approach to logistic regression, described previously, generated 8 equations (Table 3, pages 89 - 91) and 15 sets of questions ranging from 2 to 4 questions in length (Table 4 on pages 92 - 94).

All possible cut-off scores for the 15 sets of questions were analysed (refer to Table 5 on pages 95 - 97). The psychometric properties of two screening tests approach those of the 3MS (CSHA cutoff of 77/78) for the detection of mild to moderate cognitive impairment.

Table 6; Summary of sensitivity and specificity of selected derived tests vs. established screens.

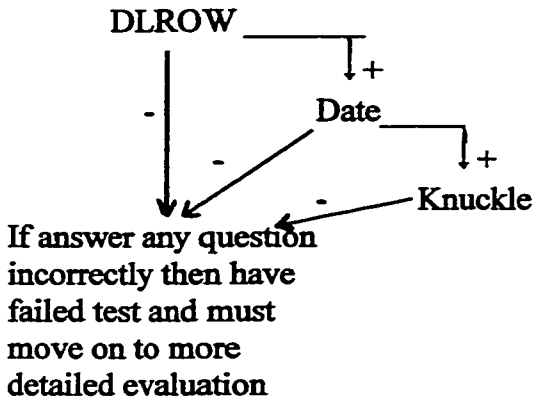
Test		Sensitivity	Specificity
3MS (cut-off 77/78)		84 %	62 %
Set 4-1 (cut-off 3/4)		**82 %**	**55 %**
Day of the week	1		
Date	1		
DLROW	1		
Year	1		
Set 3-1 (cut-off 2/3)		**80 %**	**56 %**
Day of the week	1		
Date	1		
DLROW	1		
MMSE (cut-off 25/26)		70%	73%

[5.5.2] The Recursive Partitioning approach

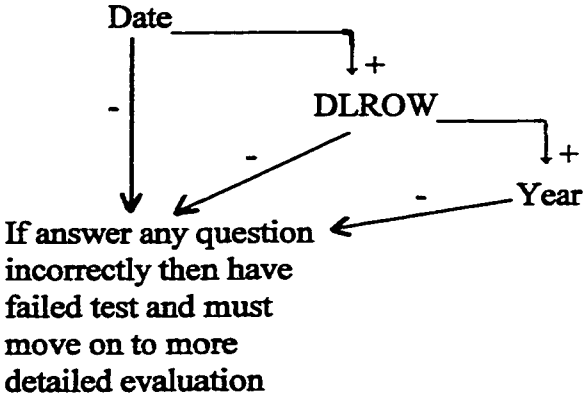
The sensitivities and specificities of the 10 basic algorithms were examined (refer to Appendix L pages 126 - 135 for details). Two algorithms approach the sensitivity of the 3MS (CSHA cutoff of 77/78) for the detection of mild to moderate cognitive impairment.

Figure 3; Sensitivity and specificity of selected cognitive screening algorithms.

	Sensitivity	Specificity
[1] Strategy 1B (or identical strategy 4B) (refer to Appendix L on page 127 for details)	**84%**	**50%**



[2] Strategy 5B (refer to Appendix L on page 135 for details)	**81 %**	**56 %**



(5.6) External validation of the "optimal" subsets of cognitive questions.

Table 8; Summary of validation results in the CSHA-2 and Ottawa General Hospital Memory Disorder Clinic (summary of Table 7 on Page 98 and Figures 4 - 6 on pages 103 - 105)

Derived screening test	*CSHA-1* (Derivation set)		*CSHA-2* (#1 validation set)			*Ottawa General MDC* (#2 validation set)	
	SENS	SPEC	* SENS	SPEC	*	SENS	SPEC
[1] set 4-1 (cut-off 3/4)	82%	55%	80%	61%		83%	60%
[2] set 3-1 (cut-off 2/3)	80%	56%	76%	62%		81%	61%
[3] strategy1B	84%	50%	77%	60%		* see below	
[4] strategy 5B	81%	51%	79%	63%		81%	62%
[5] 3MS	84%	62%					
[6] MMSE	70%	73%					

* It was not possible to validate strategy 1B in the Ottawa General Hospital Memory Disorder Clinic data because this data base did not include a 3MS question found in strategy 1B.

(5.7) ROC analysis

ROC curve analysis was performed for all logistic regression derived screens (refer to Table 9 on P. 99). The area under the curve (nonparametric calculation of AUC) for set 3-1 was .742 and for set 4-1 was .752. Both these values approach the known AUC for the MMSE (0.77) and the 3MS (0.80) in the CSHA-1 data. The usefulness of ROC curve analysis will be reviewed in greater detail. The individual ROC curves are presented in Appendix M on pages 136 - 150.

(5.8) Opinions of practising physicians

[5.8.1] The optimal length of a cognitive screening tool.

The experience of the CAGE questionnaire suggests that four questions may be at the upper limit of what busy physicians can recall and routinely employ in primary care practice. Physicians attending the University of Ottawa Geriatric Medicine Academic Rounds were asked to indicate what they felt would be the maximum number of cognitive questions which they could recall and routinely employ in clinical practice. The physicians attending were a mix of Family Medicine residents, Internists and Geriatric Medicine specialists. All of the physicians felt that they would be able to remember up to 3 questions, 70% felt they could recall 4 questions, 30% felt they could recall 5 questions and 0% felt that they could recall more than 5 questions (refer to Figure 7 on page 106). The physicians universally agreed that the use of a mnemonic or acronym would enhance recall of the questions.

A similar question was posed in the pilot survey of Élisabeth Bruyère academic Family Physicians (survey presented in Appendix K on pages 119 - 125). Very similar results were obtained in this separate group of practicing physicians with 100% responding that they could recall up to 3 questions. The percentage dropped of rapidly beyond 3 questions with 66% reporting that they could recall 4 or 5 questions, 33% felt they could recall 6 or 7 questions, and 0% beyond 7 questions (refer to Figure 7 on page 106).

[5.8.2] The results of the pilot survey of Family Physicians' opinions regarding the format of tools to screen for cognitive impairment.

The survey is presented in Appendix K (pages 119 - 125) and the results are summarized on Tables 10 and 11 (P. 100 - 101). Six of the eight academic Family Physicians practising at the Élisabeth Bruyère Family Medicine Centre responded. The results indicate that: (1) there are numerous barriers to the routine use of established tests such as the MMSE and the Clock Drawing, (2) significant disagreement exists regarding the utility of the MMSE and the Clock Drawing in routine clinical practice, (3) there exists significant discomfort with algorithmic approaches and (4) all responding physicians felt that they would be comfortable routinely employing the weighted and unweighted cognitive screens (derived from logistic regression analysis). Despite these findings, there was significant variability in the screening tests that physicians felt they would be most likely to use (Table 11 on P. 101) with some still rating the Folstein as the most likely.

The Clock Drawing deserves special mention. Overall, this was rated as the least likely to be routinely employed when a formal scoring system was required. Discussions with the responding family physicians indicate that 50% had not been aware that the clock drawing had formal scoring systems prior to this survey.

6. DISCUSSION

(6.1) Efficacy vs. Effectiveness vs. Efficiency - a question of balancing competing needs.

In previous studies researchers have generated multivariate equations which included the best questions to screen for cognitive impairment. Unfortunately these equations are too complex to routinely use in clinical practice. Cognitive screens such as the Folstein Mini-mental Status Examination do exist but are not routinely employed likely due to barriers such as time constraints [4]. This situation can be likened to the concepts of efficacy, effectiveness, and efficiency. The multivariate equations may be efficacious but lack adequate effectiveness (i.e. clinicians do not have the training or resources to use them) and efficiency (i.e. clinicians do not have the time required) to be routinely used. The published scales such as the MMSE may be efficacious and even effective (i.e. they could work in clinical practice if time permitted) but are so inefficient (yield / time) that they have not been adequately incorporated into routine clinical practice.

This study has attempted to generate simple cognitive screening tools with adequate efficacy, effectiveness and efficiency to permit and encourage routine clinical use. As one simplifies a scale, discriminative ability is usually lost. The approach employed attempts to balance the realities of busy clinical practice with the gradual loss of discriminative properties which occurs at each level of simplification of the tool. Many decisions such as which questions to select, how to simplify scoring and where to set threshold scores are inherently clinically-based judgments. As Stiell and Wells point out; "assessment of sensibility depends on judgment

rather than on statistical methods" [27]. This is a perceived weakness of this study as it may decrease the reproducibility of the analytic plan since different clinicians and researchers are unlikely to consistently make the same choices.

(6.2) The relevance of the findings of this study

The approaches employed in this study were successful in generating clinically sensible (quick, simple), accurate and readily recalled screens for cognitive impairment in community-dwelling seniors. The analytic strategies were relatively straightforward and should be readily reproducible. Consequently we not only have screens which should be acceptable to busy clinicians but we also have a methodology which will encourage the development of even better screening tools.

(6.3) The results of the derived screens in comparison to the 3MS and MMSE

The goal of this study was to develop a screening tool with a sensitivity equivalent to that of the 3MS (i.e. 84%) and then to maximize specificity. Given the fact that the 3MS was considered a suitable screening tool in the largest population based study of dementia to date (i.e. the CSHA) and also demonstrates psychometric properties superior to those of the MMSE, the goal of attempting to match the properties of the 3MS is reasonable. A sensitivity of 100% is not realistic in the field of cognitive impairment as this tends to result in extremely low specificities, such that the resulting number of false positive removes the value of screening.

In reality, clinicians will compare the psychometric properties of the new screening tool

with those of the most popular screening tool employed in primary practice - the Folstein MMSE. The MMSE did not perform well in the detection of cognitive impairment in this population. Employing the standard recommended cutoff of 25/26 resulted in a sensitivity of 70% and a specificity of 73% for the detection of mild to moderate cognitive impairment. The fact that the brief screens developed in this study surpass the sensitivity of the full MMSE should encourage their routine use.

The poor performance of the MMSE merits further discussion as it reflects limitations of the manner in which the MMSE was derived as well as the way it has been validated in previous studies. The MMSE was derived from an accumulation of clinical observations rather than being based on a conceptual model of cognitive function. Consequently it is heavily weighted on the most readily observable losses - memory and orientation. The MMSE only superficially tests other functions such as constructional apraxia (draw intersecting pentagons), agnosia (name a watch and a pen) and aphasia (repeat after me "no ifs ands or buts"). The impairment of these last three functions tends to be relatively advanced before it is detected by the MMSE.

Ideally, a test such as the MMSE should be based on a conceptual model of cognitive impairment. If such a model is not available then the test might be based on a form of cluster analysis (e.g. Factor Analysis) of a larger group of questions. The most predictive questions in each factor or cluster might be used to create the shorter test. The clustering pattern could form the basis of a model of cognitive impairment. Such an analysis is not possible with the databases available and therefore is beyond this thesis.

Despite the above, the MMSE has demonstrated better sensitivity and specificity in previous studies. Many of these studies tested the ability of the MMSE to detect dementia and not cognitive impairment. Some of these studies may have tested the MMSE on a non-representative sample of patients such as moderate to severely cognitively impaired people vs. people with normal cognition. These groups would not reflect the natural spectrum of cognitive impairment in the community. It is much easier for a test to discriminate between severely impaired people and people with normal cognition than it is to discriminate between mildly impaired people and "normals". Selecting the former situation (i.e. severely impaired vs. normal) would artificially inflate the sensitivity and specificity of the test. This is an example of spectrum bias which will be discussed in more detail below.

(6.4) Selection of optimal cut-off scores.

Post-test probabilities, positive predictive value (PPV) and negative predictive value (NPV) are all sensitive to changes in prevalence (i.e. pre-test probability in the case of cognitive impairment) which makes them difficult to employ in clinical practice. In contrast, sensitivity and specificity are relatively robust in terms of being insensitive to changes in prevalence. Consequently, sensitivity and specificity are usually employed in the selection of cut-points as well as the reporting of the psychometric properties of the derived screens. This study has focused on sensitivity and specificity for the selection of cut-off scores.

Knottnerus has stated that : " Where the line is drawn between a normal and an abnormal result is seldom objective. In general it is the result of an arbitrary decision based on a consensus

of opinion"[44]. While most of us would consider the term 'arbitrary' inappropriately strong, we would agree that in many cases cut-off points are the result of expert consensus opinion. Unfortunately this is often done implicitly rather than in an explicit, reproducible and scientific manner [38]. The statistical underpinnings of the selection of specific thresholds is frequently quite weak, a point often not well appreciated by the end user - the clinician. Many clinicians will use the published cut-offs for the MMSE with no knowledge of how these were derived, no appreciation of the limitations of each cut-off and no critical appraisal of the process of selection of cut-off scores. A number of features regarding this selection should be highlighted.

As the threshold (i.e. cut-off score or cut-point) between normal and abnormal for a screening test is raised the test becomes more difficult to pass and hence will pick up more abnormals. The test becomes more sensitive. Unfortunately this comes at a price as more "normal" subjects also fail the test (i.e. increased false positives) and the test loses specificity. Siu et al [11] clearly demonstrated this reciprocal relationship between sensitivity and specificity for the MMSE. The finding was reproduced in this study (refer to Figure 8 on P. 107). The relationship exists for most if not all other screening tests and is further demonstrated by reviewing a standard receiver operating characteristic (ROC) plot: as sensitivity increases so does 1-specificity (i.e. specificity decreases). Both sensitivity and specificity can never approach 100% simultaneously. This trade-off between sensitivity and specificity has previously been outlined by Hanley and McNeil [45]. Usually there is no perfect cutoff score. Hence, the concept of the perfect cut-off score is an illusion.

How much specificity one is willing to pay for a gain in sensitivity (or vice versa) may be based on mathematical rules but is ultimately a clinical judgment. For instance, McDowell and Kristjansson [43] have attempted to make the process of cut-point selection more explicit by applying a decision theory approach. In this technique the ratio of the posterior probabilities of false positive and false negative errors were equated with the reciprocal of the ratio of their costs or dysutilities. Cost ratios were varied from 1:10 to 10:1 and for each cost ratio a cut-point was identified that produced the smallest possible loss for each test. While this technique is a major step forward in terms of presenting an explicit methodology, the technique is still not wholly objective. The selection of which cost ratio to use, and hence which cut-point will be employed, will ultimately be a subjective clinical decision.

Kristjansson, Hill, McDowell and Lindsay [46] have also published a method for selecting cut-points based on error counting loss matrices. Once again even though the method is explicit it is not completely objective. Subjective clinical judgment must be employed to determine the relative weights of false positives and false negatives.

Perhaps the lesson here is that one cannot escape clinical judgement in applied clinical research. There are subjective components in much of clinical research. Clinical experience, expertise and judgement are important. This is not necessarily a negative point but it is important that we do openly acknowledge the subjective nature of cut-point selection.

[6.4.1] Clinical factors affecting the selection of the optimal cut-point.

As suggested above, the choice of the optimal cut-off is based on clinical judgment. This decision depends on the subjective values attached to false-positive or false-negative results. If the consequences of a false-negative result are considered to be more serious than those of a false-positive result, emphasis will be placed on cut-off thresholds with high sensitivity. This might be the situation for problems such as ankle fractures where, for legal reasons, clinicians would be unlikely to employ a decision rule if it missed any cases. Conversely, if the consequences of a false- positive test were considered to be more serious, then a cut-off resulting in a more specific test would be required. This might be true in the following scenarios; (1) serious consequences of a false-positive label as in the case of HIV status, (2) invasive follow-up tests (e.g. Surgical biopsy), or (3) invasive or potentially risky treatment for those screening positive. Once again one pays the price of more false positives as one tries to minimize the number of false negatives (and vice versa). The relative weight given to errors of misclassification (i.e. false positives or false negatives) is determined according to the intended use or anticipated effect of the test information. These weights are ultimately based on clinical judgement.

Unfortunately, it is not clear where cognitive impairment stands in this analysis. There are serious implications to missing cases of delirium or depression. As treatments become available there will be increasingly serious consequences to missing cases of early dementia (false-negatives). The costs of missing cases will continue to increase as even more effective treatments emerge. Setting thresholds which appropriately minimize these costs is a "moving

target" which will likely change over the next few decades. There are also serious consequences in terms of emotional suffering to incorrectly labeling patients with dementing illnesses such as Alzheimer's disease (i.e. false positives). Choosing the optimal level of sensitivity and specificity is not as straightforward as in the determination of the Ottawa ankle rules. It is unlikely that a single cut-off will satisfy everyone or will be appropriate in all settings.

It might be argued that in this instance we are developing a screen for cognitive impairment which includes dementia, delirium and depression. We are not labeling any patients as having specific dementing illnesses. If a patient fails this screening test they will undergo more detailed cognitive testing as well as perhaps some blood-work. There is no risk of the screen being immediately followed up with invasive or dangerous procedures. Consequently, it can also be argued that false positives are not a major concern to the individual patient. The health care system, however, would not be able to manage large numbers of false positives as Family MDs would not have the time to perform the extra confirmatory assessments and Memory Disorder Clinics already have long waiting lists. It would be reasonable to start by trying to develop a screening test with good sensitivity and at least moderate specificity. Where to precisely place the minimum acceptable sensitivity and specificity is a question which is unanswered in the literature. At the very least we should attempt to find the screen which most closely approximates the psychometric properties of the most commonly employed tests; the Folstein MMSE or the Modified Mini-mental State (3MS) examination. More detailed assessment of the selection of cut-offs and the consequences of varying the cut-off score (to rule-in or rule-out disease) would be a very valid line of future research.

[6.4.2] The role of ROC analysis

Receiver operating characteristic (ROC) curve analysis is sometimes promoted as a means for comparing, and presumably selecting among, screening or diagnostic tests (pre-conference symposium of the American Geriatrics Society 1996 annual meeting) [47]. To determine the utility of this form of analysis we will examine ROC curve analysis in greater detail.

ROC curves are plots of true positive ratios (sensitivity) vs false positive ratios (1 - specificity) for each possible cut-off score of a test. This can be regarded as a graphical display of all possible 2 X 2 contingency tables (i.e. a 2 X 2 table calculated for each possible cut-off score).

Discrimination refers to the ability of a test to separate normals from abnormals. The area under the curve (AUC) of an ROC curve estimates the overall discrimination (i.e. overall accuracy) of a test. The AUC helps determine which screening test offers the optimal combinations of sensitivity and specificity over a range of possible cut-off scores [48]. The higher the AUC the better the performance of the screening test. The AUC will fall between 0 (perfect inaccuracy) and 1 (perfect accuracy). An AUC of 0.5 indicates that the test adds no information to prediction (i.e. it is a worthless test). AUCs higher than 0.8 are considered excellent.

The AUC and ROC analysis in general may help select among several tests especially when specific cutoff scores are not determined or agreed upon. Unfortunately, ROC analysis

does not assist in the selection of these specific cutoff scores. While it is possible to set sensitivity at a particular value (e.g. 80%) and then compare curves to determine which test has the highest specificity at this preset sensitivity, it is much easier to merely scan the sensitivity vs. specificity tables as we have done in this study. In reality we will tend to select a particular screening tool because it has the most useful combination of sensitivity and specificity at a single cut-off rather than because it has a larger AUC. This is especially true in the situation where ROC curves (for different screening tools) cross. In such a case we may select the "inferior" tool (i.e. that with the lower AUC) if its performance is better at the particular sensitivity or specificity which we are targeting. Put differently, the performance within a particular region of the ROC plane may be more important than the overall AUC. The AUC may still be considered of limited use as it is important to determine that the new screening tool has a similar overall accuracy (estimated by AUC) to the MMSE or the 3MS. This would be important in situations in which one expects different cutoff scores to be employed in different clinical situations. The AUC should also approach 0.8. In the case of the screening tools developed in this study, set 3-1 (AUC = .742) and set 4-1 (AUC = .752) were close in terms of overall disciminant ability to the MMSE (AUC = .77) and the 3MS (AUC = .8).

[6.4.3] The probabilistic vs. the classification approaches.

The results of a screening tool can be presented to clinicians as the probability of the condition in question being present or as a categorization or classification (i.e. disease present or absent). These two contrasting approaches may be appropriate in different settings.

[6.4.31] The probabilistic approach

Both positive predictive value (PPV) and negative predictive value (NPV) are sensitive to changes in disease prevalence. As the prevalence increases, the PPV will increase and the NPV will decrease [48].

Unfortunately, PPV and NPV are very difficult to use in day-to-day clinical practice and have not been incorporated to any great extent in the field of cognitive impairment. Stiell and Wells [27], when discussing sensibility of decision rules, state the following; " we believe that decision rules are more likely to be used if they suggest a course of action rather than merely provide a probability of outcome.....we believe that merely giving the probability of fracture may lead to uncertainty in the clinician's decision....". Consequently, PPV and NPV may lead to clinical indecision because they do not suggest a course of action.

Another approach which has been explored in the literature is the use of likelihood ratios. Detsky et al. [37] modified the weighting scheme of the Goldman Cardiac Risk Index and then employed a nomogram (Appendix H on P. 116) to permit translation of pre-test to post-test probabilities using a "likelihood ratio line technique". While the modification likely improved the discriminant properties of the scale, this came at the price of loss of clinical practicality. It is the very rare clinician indeed who knows his local pre-test probabilities and who is also willing to use a nomogram in his everyday practice. This technique does not promote widespread routine use. Once again the post-test probabilities may add to clinical uncertainty because they do not suggest a course of action.

The counter- argument in favor of presenting the probability of disease is that this provides the clinicians with the maximum amount of information on which to base their decisions. This also permits clinicians to more accurately compare patients. This might be relevant in situations in which physicians must prioritize patients when resources are not readily available. An example would be the threshold for admission of patients with pneumonia. When there are empty ward beds physicians can be very liberal in admitting patients to hospital. However, when there is a bed shortage clinicians must be more selective in terms of which patients to admit. The decision is based on subjective clinical judgement. If a screening tool provided a probability of respiratory compromise over the next 72 hours then clinicians could use these probabilities to compare patients' risks and to prioritize patients for admission. The threshold for admission (i.e. the cut-off score) could vary according to available resources. Physicians already intuitively perform this type of analysis with varying degrees of success. Unfortunately, the threshold would still be strongly influenced by physician personality - whether they are risk takers or risk averse. Such a probabilistic approach does lead to indecision on the part of some clinicians (i.e. the risk averse) and is not popular in the practice of medicine.

[6.4.32] The classification approach

In a sense, cut-off scores are an a priori selection (often based on subjective expert consensus) of the probabilities which will be the most clinically useful and which also suggest a course of action. This does, however, limit clinician freedom and flexibility in the choice of cut-offs. The cut-offs cannot be changed when the clinical situation changes. This has been a problem for cognitive testing when the same cut-offs have been employed in different settings

(e.g. acute care, the community, nursing homes...) with very different spectra of disease. Nevertheless, preset cut-offs are welcomed by already over-taxed clinicians as these cut-offs greatly simplify their job. Unfortunately, these clinicians are often unaware of the sensitivity, specificity and limitations of the cut-off score that they are employing.

[6.4.33] Multiple cut-offs (polytomous outcomes)

This concept has been successfully incorporated into clinical care in the area of ventilation-perfusion scans which are used in the diagnosis of pulmonary emboli. The scan results are placed into one of three categories; high probability, intermediate (or indeterminate) probability, and low probability of pulmonary embolus (PE). Each of these categories suggests a different course of action. This approach is a compromise between providing the probabilities (PPV, NPV or post-test probability) at each point of measure vs. providing a single cut-off based on a priori expert consensus of where the best combination of sensitivity and specificity lies.

This study did not explore the possibility of three categories because it is unclear how this would be an improvement over two categories. Patients judged to be of high or intermediate probability of cognitive impairment would be treated the same way - they would receive further cognitive tests as well as a metabolic work-up. The categories would be collapsed and the screen results would effectively be treated as dichotomous. It is expected that such a collapsing of categories would be unlikely in other fields of medicine where the follow-up tests (e.g. angiography for PE) or treatment (e.g. anticoagulation or inferior vena cava filters for PE) carry a

risk which one would want to avoid in as many intermediate probability patients as possible. Such risky diagnostic procedures and treatments do not presently exist for cognitive impairment and consequently clinicians would be more comfortable collapsing categories. This approach does however merit further study.

[6.4.34] Matching the approach to the clinical practice

A clinician involved in a general practice such as primary care or emergency medicine must be able to diagnose and treat a large volume of patients with great variety of medical problems. Such physicians cannot become experts in each area (including cognitive impairment) and consequently would welcome expert consensus opinion on the optimal cut-off for a screening test. The classification approach is best suited to this setting.

Specialists working in settings such as memory disorder clinics have the time, training and opportunity to benefit from the increased information provided by the probabilistic approach. Positive predictive values, negative predictive values and likelihood ratios *may* have more use in the highly focused subspecialty setting. This may explain why these approaches have not become more popular in primary care and the more generalized medical fields such as emergency medicine or general internal medicine.

(6.5) Scoring - the strengths and weaknesses of weighted variables

Weighting of different variables recognizes the differential discriminability of different signs or, in this case, cognitive questions. Weighting should provide a larger selection of cut-off scores. This increased latitude in terms of selecting cut-off scores might permit finer discrimination. Surprisingly, in this study, weighting did not have a marked effect on the sensitivity and specificity of the optimal subset of questions. There was no weighted screening tool which had markedly better psychometric properties than the unweighted tools selected.

Should weighting improve the discriminant properties of a scale, then we would once again have to balance this against the ability and willingness of clinicians to use this weighted scoring in actual day to day practice. The experience with the Goldman Cardiac Risk Index has not been promising. Even those specialists (e.g. internists and cardiologists) who employ the variables in the Goldman Index rarely use the weighting system. Since one cannot employ the cut-off without using the weighting system, the cut-off score is also lost. Essentially the tool becomes a list of clinical variables which clinicians sample in their evaluation.

Furthermore, Storandt et al. [38], demonstrated that the weights may not be very stable but may be sensitive to changes in population. If the weights were found to be sensitive to even minor changes in disease spectrum then this would decrease the portability or generalizability of the cognitive screen. The area of weighting of items in a screening tool merits further study.

(6.6) Recursive Partitioning

This thesis represents the first time that recursive partitioning has been applied to the field of screening for cognitive impairment. Given the fact that 21 questions were available for analysis, there were 304,080 possible algorithms of 2,3 and 4 questions {2 X ([21 x 20] + [21 x 20 x 19] + [21 x 20 x 19 x 18]) }. Consequently, it is not feasible to study all possible algorithms and a more structured or systematic approach must be adopted.

In order to remain within the confines of objective #2 (i.e. to develop an easily reproducible methodology), only a limited number of strategies to develop recursive partitioning algorithms were explored. The criteria (maximal sensitivity, maximal specificity, maximal PPV, maximal NPV, and maximal accuracy) at each step within a single algorithm remained fixed.

While this limited and somewhat rigid approach can be justified if one wishes to develop a reproducible methodology, it may not generate the optimal series of cognitive questions. Dr. Stiell's approach to selecting each partitioning variable blends the psychometric properties of each question with clinical acumen regarding which questions clinicians will be most likely to use (personal communication). Such an approach may generate better series of questions and is the natural follow-up to the present study. Unfortunately, the approach would be very dependent on the judgement and experience of the individual developing the test and hence would not result in a reproducible methodology. Consequently, the approach did not form part of this thesis.

(6.7) Validation Studies

Often physicians will support the use of a specific clinical tool by claiming that "it has been validated". This is a relatively meaningless statement as there are several types and/or levels of validation. Furthermore, just because a validation study was performed does not imply that the tool performed well enough to justify general use. The concepts of validation are not well understood by many clinicians.

Face validity and content validity are determinations of the comprehensiveness of a tool - whether it adequately covers all the relevant domains. In terms of the screens developed in this thesis, the most obvious omission is the fact that we cannot test visuospatial function if we exclude tests involving pencil and paper.

In screening, content validity and the less formal face validity are the least important forms of validity as the aim is to collect the minimal amount of information possible for detection of the underlying disorder even if this means excluding some domains [33]. In screening for cognitive impairment the demonstration of criterion, concurrent and predictive validity are more important as they demonstrate how well the screen is assessing the dependent variable. Put differently, there appear to be two main categories of validity; (1) content (i.e. face and content validity) and (2) metric properties (concurrent, criterion and predictive validity). The former category tells us how appropriate the screen looks and the latter category tells us how well the screen works.

This thesis does not attempt to develop a screen which will predict cognitive impairment in the future. Consequently, predictive validity is not relevant. Past studies have compared new screens with the MMSE which may be considered either concurrent validity or an extremely poor form of criterion validity. Unfortunately, these studies tell us how closely the new scales predict the MMSE score. They do not tell us how well the new screens predict cognitive impairment itself. We cannot tell whether the new screens are better, worse or just as good as the MMSE. Given the poor performance of the MMSE in this data base, such concurrent validation studies are weak measures of validity. The MMSE fares so poorly that it cannot be considered a good criterion standard. The previous studies which compared new screens to the MMSE should not be considered criterion validity studies. At best they demonstrate concurrent validity.

In this thesis we have examined the criterion validity of the new screens - how well they predicted cognitive status as judged against the gold standard (i.e. the CSHA or the Ottawa General Hospital Memory Disorder Clinic standardized clinical evaluations). This is the highest form of validation. If a screen fares well at this level then all of the lower levels of validation (face, content, concurrent validity) are irrelevant. Criterion validation may be performed in a separate data base which contains similar patients and hence a similar spectrum of disease. A superior form of criterion validation would be prospective validation in a new but comparable population. Unfortunately, the resources for prospective validation are not available at this time.

(6.8) Matching analytic technique to disease (intrinsic factors) and setting (extrinsic factors)

It is possible that different analytic techniques will produce screening tools that are appropriate to different diseases, different existing clinical diagnostic strategies and different clinical settings. First we will look at different disease states in order to determine whether this might be true. This section should be considered extremely preliminary and speculative but does provide ideas for further research.

[6.8.1] Analytic technique vs. disease

Certain diagnoses such as depression are not based on a single factor but rather on a constellation of factors represented by an accumulation of criteria such as those of the DSM-IV (Diagnostic and Statistical Manual of Mental Disorders of the American Psychiatric Association). When patients have more than a threshold number of symptoms the diagnosis is made. Similar criteria exist in fields such as neurology, rheumatology, and hematology. This closely parallels the unweighted scales derived from logistic regression analysis in this thesis. It is possible that when one has multiple indicators of a single disease but when none of these indicators can be used in isolation to make the diagnosis, logistic regression analysis will be useful in deriving the best subset of indicators. Recursive partitioning has not been attempted in the field of depression. It is very possible that this analytic technique might also work although it does not match the DSM approach very well. DSM based diagnoses look at all the variables at once. This can be described as testing in parallel (i.e. all the DSM questions are asked of all patients and then analyzed together). Consequently, logistic regression (a form of testing in parallel) should be a better fit to DSM diagnoses than recursive partitioning (testing in series).

Recursive partitioning may be more suited to situations in which one has multiple factors some of which are predictive enough to make the diagnosis on their own. In the case of ankle fractures, a person with pain in the ankle so severe that they cannot bear weight probably has a fracture. These people do not need to be examined further. Their probability of a fracture is so great that they automatically merit an x-ray. There is, however, no definitive reason why logistic regression (and a weighted scale) could not also be applied to this clinical scenario.

In the analysis performed in this study the weights generated from the beta coefficients of the logistic regression equation were either one or two. There was not much variation in weights. This is consistent with the observations of psychologists that cognitive impairment (especially early or mild forms) is best detected by sampling numerous areas of cognition as many areas may be affected simultaneously. Perhaps in situations where the weights do not vary greatly, recursive partitioning will not add much to the logistic regression approach. In this study the selected recursive partitioning algorithms had sensitivities and specificities similar to those of the logistic regression derived screens. Many of the same questions were selected by both analytic strategies. Along the same line of thought, perhaps in those disease states where the logistic regression equation demonstrates very large beta coefficients (i.e. weights) for one or two variables, the recursive partitioning approach will be shown to be superior. In such situations the variables with the highest weighting may prove to be the most useful variables at the beginning of the recursive partitioning algorithm.

Recursive partitioning may also be suited to syndromes that are caused by several different diseases in two situations; (1) when the diseases are mutually exclusive (i.e. you can only have one of the diagnoses) or (2) when the diseases are not mutually exclusive but the follow-up assessment for all the diseases is the same. The latter is consistent with cognitive impairment which may be caused by dementia, delirium and / or depression (which are not mutually exclusive) and where all patients screening positive would go on to the same in-depth cognitive work-up such as a full MMSE. One might envision a recursive partitioning approach where a sensitive question for delirium is followed by a sensitive question for depression and then questions for dementia. This thesis did not explore this approach.

[6.8.2] Analytic technique vs. setting / present clinical strategy

In primary care one is working with very little initial information. Each bit of information prompts further investigation. This is the case in the assessment of cognition in the family physician's office. This process is essentially testing in series which is better modeled by recursive partitioning.

In other situations the follow-up investigations carry a significant risk and we limit these to people who have not been categorized by the safer earlier tests. This is the case in the diagnosis of pulmonary emboli - one does not perform pulmonary angiography unless the diagnosis has not been established by doppler study, ventilation-perfusion scan or spiral CT scan. This situation parallels the analytic process employed in recursive partitioning. Such risky follow-up investigations do not exist in the field of assessment of cognitive impairment.

In specialty clinics (e.g. memory disorder clinics) the specialist often has a great deal more information (i.e. cognitive tests, history, CT scan, blood work) at the starting point. This is essentially testing in parallel which is better modeled by logistic regression.

The exercise of matching analytic technique to disease and setting is a complex area which merits in-depth study by groups of seasoned clinicians and experienced researchers. As can be seen above, matching the analytic technique to the disease state is not always straightforward. Matching analytic technique to existing diagnostic strategies and clinical settings appears to be a more natural fit. The objective of this section of the thesis was merely to highlight this as a possible area for further research. This also provides some justification for pursuing the recursive partitioning approach to the assessment of cognitive impairment in primary care settings.

(6.9) The screening process in real life - targeting issues

The screens generated in this study were based on testing all available patients. In clinical practice it is unclear when family physicians would employ a cognitive screen. Discussions with practicing clinicians have uncovered an unexpected finding; there is a debate regarding whom to actually screen. Some physicians wish to apply a screen for cognitive impairment to all of their patients over the age of 65 (possibly their nurse would apply the screen). Other physicians feel that only patients they suspect of having cognitive impairment should be screened. Still others state that patients that they already feel are cognitively impaired do not need screening (they have

in effect already been screened albeit informally or intuitively). These physicians feel that it is precisely the patients over 65 years old that they are *not* concerned about who should be screened. We will analyze the logic of each of these perspectives in turn.

Screening every single elderly patient (Figure 9, scenario #1 on P. 108) demonstrates enthusiasm but does not make clinical sense. If a patient is clearly cognitively impaired (e.g. floridly delirious and hallucinating or unable to answer any questions) they do not need to be screened for cognitive impairment. They need a diagnostic work-up (i.e. step #2 in the evaluative process described in the introduction).

Screening only patients whom one suspects of being cognitively impaired (Figure 9, scenario #2 on P.108) is not supported by the literature. Numerous studies have demonstrated that physicians miss cognitive impairment in over 50% of cases. This is precisely why we are trying to develop a screen - we want to detect the 50% of cases that they miss. Once again, screening patients who are clearly cognitively impaired does not make clinical sense as was discussed in the previous paragraph.

Ideally, screens for cognitive impairment should be applied to those seniors who are not suspected of being cognitively impaired (Figure 9, scenario #3 on P. 108). This is a subset which could not be separated out for analysis in the CSHA data. Consequently, prospective validation is required. The approach depicted in Figure 9 - scenario #3 is not embraced by some family physicians. Educational strategies would be required to ensure that the screens are applied to the

correct population. It is possible that education alone may not result in appropriate selection of patients for screening. Should this be found to be true then the principles of intervening to change physician behavior [49] should be explored to optimize targeting of the screening tools.

(6.10) The opinions of practicing academic family physicians

[6.10.1] The optimal length of a cognitive screening tool.

A pattern appears to be emerging with a rapid drop-off in the number of physicians who felt that they would be able to recall a screening test once a length of 3 questions is surpassed (refer to Figure 7 on P. 106).This pattern was evident in both groups of physicians surveyed. While these findings are extremely preliminary and only 16 physicians were questioned, the results do reflect the experience of the CAGE questionnaire and, consequently, justify further research into this area. It is possible that the maximum number of acceptable cognitive questions may be increased with the use of memory enhancing strategies such as acronyms (e.g. CAGE). These are important points if one is attempting to design a screen for widespread use.

[6.10.2] The results of the pilot survey of Family Physicians' opinions regarding the format of tools to screen for cognitive impairment.

The findings depicted in Tables 10 and 11 (Pages 100 - 101) demonstrate significant disagreement regarding the utility of the MMSE and the Clock Drawing as routine screens for cognitive impairment possibly due to the identified barriers (i.e. length, need for pencil and paper, complex scoring). Despite these findings, some physicians still rated the MMSE as the most likely to be used in their clinical practice presumably due to their familiarity with the test as

well as due to the fact that it has been well validated. The clock drawing was not favored when a formal scoring system was required. As previously discussed, the scoring of the Clock Drawing is a significant barrier to widespread use and those physicians who do employ the Clock Drawing score it based on the gestalt sense that it is poor enough to warrant further testing.

The unweighted screening tool was the most acceptable. We should recognize that this does not automatically guarantee that physicians would actually employ it in practice. The most unexpected finding was that 50% of the physicians were not comfortable with the "Cognitive Screening Rules" (P. 124). While the algorithmic approach may parallel clinical work-up in the primary care setting, it is not a screening approach which is immediately acceptable to busy clinicians. Perhaps this is because in order to use an algorithm one must be able to recall not only the questions (as in the logistic regression derived scales such as set 3-1) but one must also remember the sequence of the questions and the direction of the algorithm (refer to Figures 1 and 2 on page 102). Prior to the dissemination of such algorithms, it would be advisable to explore methods (e.g. acronyms) which might improve the acceptability ratings.

(6.11) Limitations of this study

[6.11.1] Sensibility based decisions

Many decisions (e.g. selection of clinically acceptable cognitive questions, maximal number of questions per set etc.) are based on "sensibility" and hence on clinical judgment. This reduces the reproducibility of the method developed as not all clinicians or researchers will agree with the decisions made. Sensibility based decisions may be very operator dependent.

[6.11.2] Limited number of cognitive domains tested

Cognitive function can be broken down into a number of domains such as memory, language, attention, visual-spatial function etc. . When we shorten the screening test we may no longer sample certain domains. In this thesis we only employed questions with verbal responses. This does not allow us to adequately test visual-spatial function. Consequently we may end up missing some forms of cognitive impairment (e.g. delirium, mild dementia, strokes in specific locations etc.) with the shorter screen. This may explain the lack of improvement in the accuracy of the multivariate equations when one adds more than 4 questions - due to the limited selection of cognitive domains being tested one can only select from questions which represent redundant information. This merits further study. It is possible that the short screens developed compared favorably with 3MS and the MMSE because the latter two tests are themselves poor at detecting early loss in domains such as language or visual-spatial function.

[6.11.3] Spectrum bias

If the sensitivity and specificity of a screen change in populations with different disease prevalence then this is likely to be due to changes in disease spectrum rather than due to changes in prevalence alone. This form of bias, referred to as spectrum bias, was first defined by Ransohoff and Feinstein [50] and has been noted in studies which assess the generalizability of screening tests and cut-off scores to different clinical settings [51, 52] . Still other studies have noted a change in sensitivity and specificity of tests as the referral pattern changed over time (i.e. less stringent selection of candidates) in a single setting [53]. This has been called referral or selection bias but essentially creates a form of spectrum bias.

Morbidity patterns for cognitive impairment vary markedly between different sites (i.e. general office practice vs. specialist clinics vs. acute inpatient settings vs. nursing homes or chronic care hospitals). A typical example of this would be a higher prevalence of delirium in acute care. Similarly, nursing home residents will have a higher prevalence of depression and advanced dementia. Another example would be the referral of patients who are diagnostic dilemmas (e.g. very early and subtle dementias or atypical presentations) or who have very disturbing symptoms (e.g. hallucinations or behavioral problems) to specialist clinics. To further complicate matters, by the time the patients are actually seen in the specialist clinic the disease may have progressed to the point where it is much more readily diagnosed than at the time they were initially referred from the primary care setting. Whether the specialists see primarily very early and subtle cases or more advanced cases will vary between specialist practices (practice-to-practice and regional variation).

This creates a form of spectrum bias (a.k.a. case- mix bias or referral bias) which will significantly affect the operating characteristics of a screening test. For example a test may perform very well (high sensitivity and specificity) in separating severely cognitively impaired people from others who are completely cognitively normal. The spread of the test scores for these two groups may not even overlap. Unfortunately, the same screening test may fail to adequately discriminate mildly impaired people from normals because the spread of scores for the two groups overlaps significantly in the new setting. This situation is commonly seen when tests are first derived in inpatient settings or specialist clinics where severe cases are more prevalent and then are applied without re-validation to the community setting where milder cases are more common. Appendix N (P. 151)[46] demonstrates the significant overlap of 3MS scores between demented, CIND (Cognitive Impairment Not Demented) and normals in the CSHA-1. Where such an overlap exists between diagnostic groups, sensitivity and specificity will be limited. As the cut-off score is raised, more people with dementia will fail the test and hence it will become more sensitive. Unfortunately, since the distributions of scores for the CIND and Normal group overlap those of the demented group, more Normals will also fail the test and the specificity will fall. This overlap of scores explains why we cannot approach 100% sensitivity and 100% specificity simultaneously as we vary cut-off scores as depicted in Figure 8 (P. 107).

As a result of the above one cannot take a screen and an optimal cut-off score developed in one setting and freely apply it to another setting. This is often not appreciated in the published literature and in the application of new diagnostic tests. This often results in new diagnostic tests failing to live up to their initial promise. What would be required is either redeveloping the tool

in the new setting or at the very least prospective validation and adjustment of cut-off scores.

The application of cognitive screening tools derived in this study to other settings such as acute care, nursing homes or specialist clinics should be done with extreme caution as the spectrum of disease is expected to be very different in these settings. Re-derivation or at least re-validation of the screen (which likely will involve a change in cut-off scores) in these new settings is not only recommended but should be considered an absolute requirement.

The major limitation of this data base research is that we cannot anticipate which patients Family Physicians will target for the use of these screens and, consequently, we cannot determine the definitive sensitivities and specificities of the screening tools which have been developed. In clinical practice we do *not* need a screen if we already suspect that a patient is suffering from cognitive impairment. Unfortunately, the literature demonstrates that 50% of cases of cognitive impairment will be missed if we rely on clinical suspicion alone. It is precisely the patients whom we do *not* suspect of cognitive impairment (i.e. which include the 50% of cases we missed) who need to be screened. This is depicted in Figure 9, scenario #3 (P.108). This is a subset of the original CSHA patients and thus creates a form of spectrum bias.

The targeting of patients for such screening tools is likely much more complex. Some physicians will not be comfortable following scenario #3 and will likely continue to apply the screens to patients they strongly suspect of being cognitively impaired. Conversely, some patients with mild impairment may retain insight into their evolving deficits and may request

further evaluation. Such patients likely merit more detailed assessment rather than the application of the screen. In this thesis all patients with severe cognitive impairment (i.e. MMSE < 11) were removed from analysis because it was felt that in general such patients should be readily identifiable and do not require screening. This may be an oversimplification. Some patients with relatively advanced dementia may be missed if they are quiet and cooperative. The screens may actually benefit this group.

As a consequence of the above, it is entirely unpredictable who will be screened in the primary care setting as well as who would benefit from screening. Not only can we not predict actual targeting, but we also cannot define optimal targeting. This data base research cannot provide definitive sensitivities and specificities without such targeting information. Prospective validation in the primary care setting is required.

[6.11.4] Work-up bias (a.k.a. verification bias or sequential-ordering bias)

In community studies it is often not feasible to fully evaluate all participants due to limitations in resources (i.e. time, money, trained personnel etc.). If all the subjects who screen positive but only a subset of the majority who screen negative are referred for full diagnostic work-up, this will inflate sensitivity, decrease specificity and hence will change the ROC curves and the AUCs. This effect is called verification or work-up bias [46]. This is precisely what happened in the CSHA. Lack of correction of verification bias is a minor limitation at most.

This form of bias remains uncorrected in most general population studies. Formulae for the correction of verification bias have been published [54-56, 59, 60]. Unfortunately, it is unclear how well these approaches truly correct for verification bias due to the following concerns :

[1] The corrected sensitivities and specificities differ depending on the particular formula employed. (Corrected sensitivity can differ by as much as 28%, and corrected specificity by up to 10%).

	Set 3-1		3MS (77/78 cut-off)	
- Naive (uncorrected) values				
- $sensitivity_{naive}$	.8		.84	
- $specificity_{naive}$		.56		.62
- Knotternus method [60]				
- $sensitivity_{corrected}$	.285		.344	
- $specificity_{corrected}$		.927		.942
- Choi method [54]				
- $sensitivity_{corrected}$	.33		.166	
- $specificity_{corrected}$		1.05		.978
- Greenes & Begg method [59]				
- $sensitivity_{corrected}$	.377		.45	
- $specificity_{corrected}$		.897		.912

[2] The thresholds set by these correction formulae are either exceedingly high or do not correspond to thresholds one would select in preliminary data analysis. For instance, if only 10% of test negative patients go on for further testing and one were aiming for a corrected sensitivity = 80% and corrected specificity = 60% (as in this thesis), then

according to the Knotternus method one would have to select screening tools with a naive sensitivity = 98% and a naive specificity = 13%.

The formulae suggest that the corrected sensitivity for the 3MS is between 16 - 34%! This is drastically lower than previously published sensitivities of the 3MS. Such extreme "corrections" would tend to invalidate most if not all existing cognitive screens.

These findings suggest that the formulae may be somewhat excessive in the degree to which they correct. The literature supports the direction of correction (i.e. decreasing sensitivity and increasing specificity) but no studies were found which validated the degree of the correction.

In this thesis we are comparing a number of screening tests which are presumably equally affected by verification bias. If this form of bias remains constant then the effect on the derived screening tests and on the 3MS should be similar. Consequently, our selection of the screening test which most closely matches the 3MS should be unaffected.

If we examine the effect of the formulae employed to correct for verification bias, we note that they appear to have a similar "corrective" effect on all the screens being considered.

	Set2-1		Set 3-1		Set 4-1		3MS	
- Naive (uncorrected) values								
- sensitivity$_{naive}$	*.12*		*.8*		*.82*		*.84*	
- specificity$_{naive}$		.99		.56		.55		.62
- **ORDER**	**1**		**2**		**3**		**4**	
- Knotternus method [60]								
- sensitivity$_{corrected}$	*.01*		*.285*		*.31*		*.344*	
- specificity$_{corrected}$		.999		.927		.924		.942
- **ORDER**	**1**		**2**		**3**		**4**	

If one is attempting to rank and select tests based on how close their sensitivities are to that of the 3MS (as in this thesis), then the rank order appears to be unaffected by the verification bias correction formulae. If we are attempting to match the psychometric properties of the published cut-offs of the 3MS as closely as possible (as indicated on page 31) then the correction formulae do not change this relationship. The psychometric properties of the 3MS change to the same degree as the derived screens. The screens whose properties are closest to those of the 3MS should still be the closest after correction.

Given; (1) the major shifts in spectrum of disease expected with the recommended change in selection of patients for screening (i.e. change in referral bias) , (2) the unpredictability of these spectrum shifts, (3) the fact that the formulae correcting for verification bias do *not* correct for the change in referral bias, (4) the uncertainty regarding how well the above formulae correct for verification bias (i.e. variability of corrected results, uncertainty regarding which formula to use, lack of validation of the formulae themselves, the fact that they would invalidate all existing cognitive screens etc.), the old adage still holds true that *the validity of a test is best determined in a population of the same type as that to which it will be applied.* The bottom line is that prospective validation in the proper population is required before sensitivities and specificities can be declared definitive.

The fact that correction for verification bias does not change the relationship of derived screens to the 3MS further supports the contention that the lack of correction for this bias is a minor limitation of this thesis.

Despite the above, we should attempt to anticipate the effect of verification bias. Future validation studies should not focus exclusively on the selected screens. If feasible (i.e. if resources and physician compliance permit) screens with higher sensitivity and lower specificity should be studied concurrently as they may correct to a better sensitivity and specificity than the proposed screening tools.

(6.12) Strengths of this study

This project differs from studies previously published [28-32] in a number of aspects. This thesis focused on cognitive impairment rather than on dementia. As previously discussed, screening for cognitive impairment is more clinically relevant.

In order to develop a screening test which is useful in clinical practice, three modifications of prior logistic regression analyses [30-32] have been made: (1) preselection of variables (cognitive questions) based on their anticipated acceptability to clinicians; (2) use of a sequential weighting scheme to improve the selection of sensitive or specific cognitive questions; and (3) re-scoring of the variables selected by the logistic regression algorithm. The sensitivity and specificity of the first four variables was determined "outside" of the regression equation in a more clinically usable format (i.e. simple integer scoring systems similar to that of the CAGE questionnaire[34] or the Goldman Cardiac Risk Index[36]).

Recursive partitioning was employed to determine if this form of analysis can generate algorithmic screening tests similar to the Ottawa ankle rules. This technique has never been applied to the field of cognitive screening research.

Past studies have selected variables from the Folstein MMSE (30 point scale). In this study the number of variables (cognitive questions) analyzed was increased by employing the larger 100 point Modified Mini-mental State Examination (3MS) which contains the MMSE. This should increase the number of potential sets of questions which can be generated.

This study employed full cognitive evaluation by a multidisciplinary team with expertise in cognitive disorders as the criterion standard. This is a much better "Gold Standard" than the overall MMSE scores employed as the criterion standard in the prior studies.

(6.13) Future Directions / Analyses

The above analysis should be considered preliminary and exploratory. The exercise has yielded many ideas for follow-up research. This thesis will serve as the foundation or blueprint for a long-term stream of research with many practical clinical applications. Real-time clinical validation of the derived screening tools within the primary care setting would be the priority. Numerous other potential approaches are reviewed in Appendix O (pages 152 - 155).

7. CONCLUSION

(7.1) The optimal screening tool(s) generated by this study (objective #1)

This thesis has generated cognitive screening tools and has narrowed down the number of screens which merit prospective validation. Given the limitations regarding how much can realistically be studied in the busy primary care setting, these results will be very useful in increasing the feasibility of such prospective research.

This study validated two logistic regression based screens (one 3 question screen, and one 4 question screen) and two 3 question recursive partitioning algorithms with sensitivities and specificities between those of the MMSE (70% sensitivity, 73% specificity) and the 3MS (84% sensitivity, 62% specificity) for the detection of mild to moderate cognitive impairment.

Based on the psychometric properties previously calculated and the responses to the survey of family physicians, the optimal set of cognitive questions appears to be set 3 -1:

Set 3-1 (cut-off 2/3) Sensitivity = 80 %, Specificity = 56 %

Day of the week	1
Date	1
DLROW	1

This will be known as the *Ottawa 3D screen* for cognitive impairment to assist in the recall of the 3 questions.

When one includes the other three tests with similar psychometric properties and those tests with higher sensitivity and lower specificity (to anticipate verification bias effect), one is left with two single questions, three logistic regression based sets and three recursive partitioning based algorithms (refer to Appendix P on pages 156 - 158). Screening tools for prospective validation should be selected from these potential candidates.

(7.2) Recommended analytic strategy for the development of clinically practical screens for cognitive impairment (objective #2):

The following steps are recommended to guide future researchers attempting to develop brief, clinically sensible screens for cognitive impairment. These should be considered in conjunction with the checklist of standards for the 6 stages in the development of a clinical decision rules [27] (Appendix Q on P. 159).

[Step 1] Selection of appropriate data bases

This incorporates the concept of spectrum of disease. Put simply, the patients in the derivation and validation data sets should match those for whom the screening tool is intended. The application of screens and cut-off scores developed in one setting (e.g. specialty clinics or acute care) to another setting (e.g. nursing homes or the community) without proper validation in the new settings is inappropriate.

[step #2] Selection of appropriate criterion standards

Previous studies have selected screening questions from established screens such as the MMSE and have then used the MMSE as the criterion standard. This is considered unacceptable. The recommended criterion standard for research is a standardized clinical examination which incorporates diagnostic information from the medical history, the physical examination, and appropriate investigations (the latter may be based on consensus recommendations [58]). The diagnostic procedures employed in the Canadian Study of Health and Aging may be considered a model.

[step #3] Selection, Coding and Scoring of Screening Questions

Based on the concept of clinical sensibility, the cognitive questions derived from an established longer cognitive test should be selected and scored in the fashion which maximizes acceptability to busy clinicians. Examples might include avoidance of questions that require "props" (i.e. pencil, paper, cue cards...) and dichotomization of all individual question scores in a manner which will maximize clinician recall (of the scoring system).

[step #4] Selection of the multivariate analytic technique

Given the dichotomous dependent variable (cognitive impairment vs. normal), the recommended dichotomization of independent variables (i.e. cognitive questions) and the fact that cognitive scores are known to have non-normal distributions even prior to dichotomization of responses, logistic regression is the most appropriate multivariate technique. Weighting of variables utilizing beta coefficients is suggested but the gains must be carefully weighed against

the difficulty clinicians will have in recalling and using such weights.

Where investigator experience permits, it is recommended that other analytic techniques such as recursive partitioning and cluster analysis also be explored.

[step #5] Translating the results of multivariate analyses into a clinically acceptable and clinically usable format.

Multivariate equations are not useful in the clinical setting and hence it is strongly recommended that researchers do not stop at this point. The equations should be translated into simple scales by selecting the first 2,3 or 4 variables which enter the equation. These are the most predictive variables. Recognizing the law of diminishing returns (i.e. the later a variable enters the equation, the less predictive it is), the number of questions should only be lengthened beyond 4 if this adds significantly to the discriminant properties of the scale.

Similarly, weighting of variables (via the logistic regression beta coefficients) should only be employed if it substantially improves sensitivity and/or specificity.

[step #6] Setting cut-off (threshold) scores

Once the simplified scales have been established, then all possible cut-off scores should be explored. The sensitivity and specificity of each cut-off score should be clearly demonstrated. The rationale for selecting a specific cut-off should be explicitly stated.

[step #7] Validation

All forms of multivariate analysis tend to overfit their results to the available data. Consequently, screening tests derived from such analyses often do not operate as well in their target clinical populations as they did in the original derivation data. Prior to use in clinical practice, screening tests must be validated in a different data base or ideally should be validated prospectively. Cut-off scores and weights may require revision after such validation. Definitive sensitivities and specificities must be derived from the results of these validation studies.

[step #8] Acceptability studies

The above steps should result in the development of valid and clinically practical screening tools. This does not, however, guarantee that such a tool will be adopted into clinical practice. It is recommended that the opinion of the intended users (i.e. primary care physicians in this case) be elicited. This may further enhance the clinical utility of the final product.

Further steps in the development of screening tools may derived from Stiell and Wells' checklist for the 6 stages in the development of a clinical decision rule (Appendix Q on page 159). The underlined steps in the checklist have been addressed in this thesis.

****

8. REFERENCES

1. Rockwood K. Forgetfulness: is dementia the cause ?
   The Canadian Journal for CME 1995; June: 51 - 57

2. Ritchie K, Fuhrer R. A comparative study of the performance of screening tests for senile dementia using Receiver Operating Characteristics Analysis. J Clin Epidemiol 1992; 45(6): 627 - 637

3. Solomon PR, Hirschoff A, Kelly B et al. A seven minute neurocognitive screening battery highly sensitive to Alzheimer's Disease. Arch Neurol 1998; March: 349 - 355.

4. Bush C, Kozak K, Elmsie T. Screening for cognitive impairment in the elderly.
   Can Family Physician 1997; 43:1763 - 1768

5. Illife S, Booroff A, Gallivan S et al. A screening for cognitive impairment in the elderly using the Mini-Mental State Examination. Br J Gen Pract 1990; 40: 277 - 279.

6. O'Connor DW, Pollitt PA, Hyde JB, et al. Do general practitioners miss dementia in elderly patients? BMJ 1988; 297:1107 - 1110.

7. Bowers J, Jorm AF, Henderson S, Harris P. General practitioners' detection of depression and dementia in elderly patients. Med J Aus 1990; 153: 192 - 196.

8. McCartney JR, Palmateer LM. Assessment of cognitive deficit in geriatric patients.
   JAGS 1985; 33: 483 - 488.

9. Somerfield MR, Weisman CS, Ury W, et al. Physician practices in the diagnosis of dementing disorders. J Am Geriatr Soc 1991; 39: 172 - 175.

10. Parboosingh EJ. Screening and diagnosis of dementia in family practice including the role of the geriatrician. Poster presented at the Annual Meeting of the Royal College of Physicians and Surgeons of Canada; 1992 Sept 11 - 14 Ottawa.

11. Siu AL. Screening for dementia and investigating its causes.
    Annals of Internal Medicine 1991; 115(2): 122 - 132.

12. Feinstein AR. Clinimetrics. New Haven, CT: Yale University Press, 1987.

13. Stiell IG, Greenberg GH, McKnight RD et al. A study to develop clinical decision rules for the use of radiography in acute ankle injuries. Annals of Emergency Medicine 1992; 21(4): 384 - 390.

14. Stiell IG, Greenberg GH, McKnight RD et al. Decision rules for the use of radiography in acute ankle injuries - refinement and prospective validation. JAMA 1993; 269 (9): 1127 - 1132.

15. Canadian Study of Health and Aging Working Group. Canadian Study of Health and Aging study methods and prevalence of dementia. Can Med Assoc J 1994;150(6): 899 - 912.

16. Graham JE, Rockwood K, Beattie BL et al. Prevalence and severity of cognitive impairment with or without dementia. Lancet 1997; 349:1793 - 1796.

17. Periodic Health Examination, 1991 update; 1. Screening for cognitive impairment in the elderly. Can Med Assoc J 1991; 144(4): 425 - 431

18. Rogers SL, Friedhoff LT, and the Donepezil Study Group. The efficacy and safety of donepezil in patients with Alzheimer's Disease: results of a US multicentre, randomized, double-blind, placebo-controlled trial. Dementia 1996; 7: 293 - 303.

19. Rogers SL, Farlow RS, Doody RS, et al. A 24-week double-blind, placebo-controlled trial of donepezil in patients with Alzheimer's Disease. Neurology 1998; 50: 136 - 145.

20. Rogers S, Friedhoff L . Long-term efficacy and safety of donepezil in the treatment of Alzheimers disease; an interim analysis of the results of a US multicentre open label extension study. European Neuropsychopharmacology 1998; 8: 67 - 75.

21. Burback D, Molnar F, St. John P, Man-Son-Hing M. Key methodological features of randomized controlled trials of Alzheimer's Disease therapy. Dementia 1999; 10; 534 - 540

22. Burns A, Lawlor B, Craig S. Assessment Scales in Old Age Psychiatry. Martn Dunitz Ltd 1999

23. McDowell I, Newell C. Measuring Health (second edition) New York, Oxford Press, 1996. pp.287 - 384.

24. Froehlich TE, Robison JT, Inouye SK. Screening for dementia in the outpatient setting: the Time and Change test. JAGS 1998; 46: 1506 - 1511.

25. Folstein MF, Folstein SE, McHugh PR. "Mini-Mental State" A practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res 1975; 12: 189 - 198.

26. Tombaugh TN, McIntyre NJ. The Mini-Mental State Examination : A comprehensive review. JAGS 1992; 40: 922 - 935.

27. Stiell IG, Wells GA, Methodologic standards for the development of clinical decision rules in emergency medicine. Annals of Emergency Medicine 1999; 33(4): 437 - 447.

28. Paveza GJ, Cohen D, Blaser CJ, Hagopian M.. A brief form of the Mini-mental State Examination for use in community care settings. Behavior, Health and Aging 1990; 1(2): 133 - 139.

29. Koenig HG. An abbreviated Mini-mental State Exam for medically ill older adults. JAGS 1996; 44(2): 215 - 216.

30. Klein LE, Roca RP, McArthur J et al. Diagnosing Dementia - univariate and multivariate analyses of the Mental Status Examination. JAGS 1985; 33(7): 483 - 488.

31. Magaziner J, Bassett SS, Hebel JR. Predicting performance on the Mini-mental State Examination ; use of age- and education-specific equations. JAGS 1987; 35(11): 996 - 1000.

32. Braekhus A, Laake K, Engedal K. The Mini-mental State Examination : identifying the most efficient variables for detecting cognitive impairment in the elderly. JAGS 1992; 40(11): 1139 - 1145.

33. Ritchie K. The screening of cognitive impairment in the elderly : a critical review of current methods. J Clin Epidemiol 1988; 41(7): 635 - 643

34. Buchsbaum DG, Buchanan RG, Centor RM et al. Screening for alcohol abuse using scores and likelihood ratios. Annals of Internal Medicine 1991; 115(10): 774 - 777.

35. Wenrich MD, Paauw DS, Carline JD et al. Do primary care physicians screen patients about alcohol intake using the CAGE questions? J Gen Intern Med 1995; 10: 631 - 634.

36. Goldman l, Caldera DL, Nussbaum SR et al. Multifactorial index of cardiac risk in non-cardiac surgical procedures. N Engl J Med 1977; 297: 845 - 850.

37. Detsky AS, Abrams HB, McLaughlin JR et al. Predicting cardiac complications in patients undergoing non-cardiac surgery. J Gen Intern Med 1986; 1: 211 - 219

38. Storandt M, Botwinick J, Dannziger WL et al. Psychometric differentiation of mild senile dementia of the Alzheimer type. Arch Neurol 1984; 41: 497 - 499..

39. Eslinger PJ, Damasio AR, Benton AL, Van Allen AM. Neuropsychologic detection of abnormal mental decline in older persons. JAMA 1985; 253: 670 - 674.

40. Hosmer DW, Lemeshow S. Applied Logistic Regression. John Wiley & Sons 1989.

41. Ciampi A, Chang C-H, Hogg SA et al. Recursive partitioning: a versatile method for exploratory data analysis in biostatistics. Biostatistics 1987; 23 - 50.

42. Teng EL, Chui HC. The Modified Mini-mental State (3MS) Examination. J Clin Psychiatry 1987; 48: 314 - 318.

43. McDowell I, Kristjansson B, Hill GB, Hébert R. Community screening for dementia: The Mini-mental State Exam (MMSE) and Modified Mini-mental State Exam (3MS) compared. J Clin Epidemiol 1992; 45: 581 - 586.

44. Knottnerus JA. Interpretation of diagnostic data: an unexplored field in general practice. Journal of the Royal College of General Practitioners 1985 ;35 : 270 - 274.

45. Hanley JA, McNeil BJ. The meaning and use of the area under a Receiver Operating Characteristic (ROC) curve. Radiology 1982; 143: 29 - 36.

46. Kristjansson B, Hill G, McDowell I, Lindsay J. Optimal cut-points when screening for more than one disease state: an example from the Canadian Study of Health and Aging. J Clin Epidemiol 1996; 12: 1423 - 1428.

47. Bush K, Kivlahan DR, McDonell MB et al. The AUDIT Alcohol Consumption Questions (AUDIT - C) ; an effective brief screening test for problem drinking. Arch Intern Med 1998; 158: 1789 - 1795.

48. Baldessarini RJ, Finklestein S, Arana GW. The predictive power of diagnostic tests and the effect of prevalence of illness. Arch Gen Psychiatry 1983; 40: 569 - 573.

49. Poses RM. One size does not fit all: questions to answer before intervening to change physician behaviour. Journal on Quality Improvement 1999; 25(9): 486 - 495.

50. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. N Engl J Med 1978; 299: 926 - 930.

51. Lachs MS, Nachamkin I, Edelstein PH et al. Spectrum bias in the evaluation of diagnostic tests : lessons from the rapid dipstick test for urinary tract infection. Annals of Internal Medicine 1992; 117: 135 - 140.

52. Knottnerus JA. The effects of disease verification and referral on the relationship between symptoms and disease. Med Decis Making 1987; 7: 139 - 148.

53. Rozanski A, Diamond GA, Berman D et al. The declining specificity of exercise radionuclide ventriculography. N Engl J Med 1983; 309: 518 - 522.

54. Choi BK. Sensitivity and specificity of a single diagnostic test in the presence of work-up bias. J Clin Epidemiol 1992; 45: 581 - 586.

55. Gray R. Begg CB, Greenes RA. Construction of Receiver Operating Characteristic Curves when disease verification is subject to verification bias. Med Decis Making 1984; 4: 151 - 164.

56. Hunnink MGM, Richardson DK, Doubilet PM, Begg CB. Testing for fetal pulmonary maturity : ROC analysis involving covariates, verification bias, and combination testing. Med Decis Making 1990; 10: 201 - 211.

57. Huppert FA, Tym E. Clinical and neuropsychological assessment of dementia. Br Med Bull 1986; 42: 11 - 18

58. Patterson C, Gauthier S, Bergman H et al. Management of Dementing Disorders; conclusions from the Canadian Consensus Conference on Dementia. CMAJ 1999; 160 (12 Suppl): s1-s15.

59. Greenes RA, Begg CB. Assessment of Diagnostic Technologies; Methodology for Unbiased Estimation from Samples of Selectively Verified Patients. Invest Radiol 1985; 20: 751-756.

60. Knottnerus JA, Leffers P. The Influence of Referral Patterns on the Characteristics of Diagnostic Tests. J Clin Epidemiol 1992; 45(10): 1143 - 1154.

Tables & Figures

TABLE 2; Results of bivariate (χ^2) analysis

Question	χ^2	P-value	Sens(%)[95% CI]	Spec(%)[95% CI]
1. In which year were you born?	23.09	0.0000	8 [6, 11]	98 [96, 99]
2. What is your birth date?	14.60	0.0001	6 [4, 8]	99 [97, 99]
3. In which month were you born?	12.93	0.0003	3 [2, 5]	99 [99, 100]
4. Three word registration.	26.53	0.0000	17 [14, 20]	93 [91, 95]
5. Count from 5 to 1 backwards.	105.1	0.0000	61 [57, 65]	69 [66, 73]
6. Spell WORLD backwards.	106.2	0.0000	61 [56, 65]	70 [66, 73]
7. Three word recall	47.19	0.0000	95 [93, 97]	18 [15, 22]
8. What is today's date?	111.1	0.0000	53 [48, 57]	78 [74, 81]
9. Which month are we in?	72.89	0.0000	15 [12, 18]	99 [97,99]
10. What year is it?	103.5	0.0000	23 [19, 26]	97 [95, 98]
11. What season are we in?	31.76	0.0000	12 [9, 15]	97 [95, 98]
12. What day of the week is it?	123.0	0.0000	24 [20, 28]	98 [96, 99]
13. What province are we in?	28.59	0.0000	6 [5, 9]	99 [98, 100]
14. What country are we in?	18.78	0.0000	13 [10, 16]	95 [93, 96]
15. What city are we in?	16.56	0.0000	6 [4, 8]	98 [97, 99]
16. What building are we in? *	0.001	0.9224	1 [0, 2]	99 [98, 100]
17. What is this (forehead)? *	1.44	0.2300	4 [3, 6]	97 [95, 98]
18. What is this (chin)?	8.704	0.0032	3 [2, 4]	99 [99, 100]
19. What is this (shoulder)? *	3.212	0.0731	4 [2, 6]	98 [97, 99]
20. What is this (Elbow)?	4.828	0.0280	4 [2, 6]	98 [97, 99]
21. What is this (Knuckle)?	18.69	0.0000	23 [19, 27]	87 [84, 90]
22. Similarity (arm/leg)	33.15	0.0000	93 [92, 96]	17 [14, 20]
23. Similarity (laughing/crying)	61.78	0.0000	86 [83, 89]	34 [30, 38]
24. Similarity (eating/sleeping)	37.72	0.0000	88 [84, 90]	27 [24, 31]

* Not entered into the multivariate analysis due to their lack on significance on bivariate analysis.

TABLE 3; Logistic regression equations

	β	S.E.	Wald	df	Sig	R	Exp β
[1] Equation 1 (source of question sets 4-1 & 4-3)							
What is the day of the week?	1.85	.3188	33.81	1	0.000	.1425	6.39
Spell WORLD backwards	1.05	.1367	58.95	1	0.000	.1906	2.86
What is the date?	0.80	.1461	29.86	1	0.000	.1333	2.22
What is the year?	1.51	.2916	26.87	1	0.000	.1259	4.53
Constant	-4.02	.3979	102.48	1	0.000		
Sensitivity = .50, Specificity = .86 Accuracy = .69							
[2] Equation 2 (source of question sets 4-2 & 4-4)							
What is the day of the week?	2.19	.3133	49.06	1	0.000	.1733	8.98
Spell WORLD backwards	.98	.1375	50.31	1	0.000	.1755	2.65
What is the date?	.79	.1458	29.56	1	0.000	.1326	2.21
What is the difference between laughing and crying?	.85	.1705	24.79	1	0.000	.1206	2.34
Constant	-3.10	.3130	98.25	1	0.000		
Sensitivity = .66, Specificity = .73 Accuracy = .70							

TABLE 3: Logistic regression equations (cont)

	β	S.E.	Wald	df	Sig	R	Exp β
[3] Equation 3 (source of question sets 3-1 & 3-4)							
What is the day of the week?	2.16	.3103	48.29	1	0.000	.1718	8.64
Spell World backwards	1.10	.1346	66.65	1	0.000	.2031	3.00
What is the date?	.87	.1439	36.21	1	0.000	.1477	2.38
Constant	-2.99	.3090	93.67	1	0.000		
Sensitivity = .45, Specificity = .89 Accuracy = .68							
[4] Equation 4 (source of question sets 3-2 & 3-5)							
What is the day of the week?	2.17	.3132	48.02	1	0.000	.1713	8.76
Spell World backwards	1.14	.1343	71.58	1	0.000	.2107	3.11
What is the year?	1.62	.2876	31.67	1	0.000	.1376	5.04
Constant	-3.94	.3928	100.8	1	0.000		
Sensitivity = .72, Specificity = .67 Accuracy = .70							
[5] Equation 5 (source of question sets 3-3 & 3-6)							
What is the day of the week?	1.87	.3140	35.33	1	0.000	.1458	6.47
What is the year?	1.64	.2850	32.91	1	0.000	.1404	5.13
What is the date?	.93	.1411	43.19	1	0.000	.1621	2.53
Constant	-3.66	.3862	89.59	1	0.000		
Sensitivity = .59, Specificity = .74 Accuracy = .67							

TABLE 3; Logistic regression equations (cont.)

	β	S.E.	Wald	df	Sig	R	Exp β
[6] Equation 6 (source of question sets 2-1,2-4 & 2-5)							
What is the day of the week?	2.55	.3034	70.69	1	0.000	.2093	12.82
Spell WORLD backwards	1.20	.1319	82.54	1	0.000	.2267	3.31
Constant	-2.85	.3059	86.59	1	0.000		
Sensitivity = .69, Specificity = .69 Accuracy = .69							
[7] Equation 7 (source of question set 2-2)							
What is the day of the week?	2.2	.3071	51.62	1	0.000	.1779	9.08
What is the year?	1.77	.2801	40.12	1	0.000	.1559	5.90
Constant	-3.49	.3773	85.50	1	0.000		
Sensitivity = .34, Specificity = .95 Accuracy = .66							
[8] Equation 8 (Source of question set 2-3)							
Spell WORLD backwards	1.13	.1305	74.52	1	0.000	.2151	3.01
What is the date?	1.20	.1356	78.52	1	0.000	.2209	3.32
Constant	-1.26	.1275	97.17	1	0.000		
Sensitivity = .79, Specificity = .57 Accuracy = .68							

TABLE 4 ; The unique sets of questions derived from logistic regression analysis

Coding of the sets will be as follows; the first number will refer to the number of questions and the second number will refer to the specific set of this length. For example the third set of 4 questions will be labelled 'set 4-3'.

Sets of 4 questions	Questions	Weights
Set 4-1	What is the day of the week?	1
	Spell WORLD backwards	1
	What is the date?	1
	What is the year?	1
Set 4-2	What is the day of the week?	1
	Spell WORLD backwards	1
	What is the date?	1
	What is the difference between laughing and crying?	1
Set 4-3 (weighted)	What is the day of the week?	2
	Spell WORLD backwards	1
	What is the date?	1
	What is the year?	2
Set 4-4 (weighted)	What is the day of the week?	2
	Spell WORLD backwards	1
	What is the date?	1
	What is the difference between laughing and crying?	1

(Table 4 continued)

Sets of 3 questions	Questions	Weights
Set 3-1	What is the day of the week?	1
	Spell World backwards	1
	What is the date?	1
Set 3-2	What is the day of the week?	1
	Spell World backwards	1
	What is the year?	1
Set 3-3	What is the day of the week?	1
	What is the year?	1
	What is the date?	1
Set 3-4 (weighted)	What is the day of the week?	2
	Spell World backwards	1
	What is the date?	1
Set 3-5 (weighted)	What is the day of the week?	2
	Spell World backwards	1
	What is the year?	2
Set 3-6 (weighted)	What is the day of the week?	2
	What is the year?	2
	What is the date?	1

(Table 4 continued)

Sets of 2 questions	Questions	Weights
Set 2-1	What is the day of the week?	1
	Spell WORLD backwards	1
Set 2-2	What is the day of the week?	1
	What is the year?	1
Set 2-3	Spell WORLD backwards	1
	What is the date?	1
Set 2-4 (weighted)	What is the day of the week?	2
	Spell WORLD backwards	1
Set 2-5 (weighted)	What is the day of the week?	3
	Spell WORLD backwards	1

TABLE 5; Sensitivities and Specificities of various cut-points **(abnormal / normal)**

Sets of 4 questions	Cut-point	Sensitivity (%) [95% CI]	Specificity (%)[95% CI]
Set 4-1	0/1	9 [7, 11]	100 [99, 100]
	1/2	22 [18, 25]	99 [98, 100]
	2/3	48 [44, 52]	88 [85, 90]
	3/4	**82 [78, 85]**	**55 [51, 59]**
Set 4-2	0/1	12 [9, 15]	99 [98, 100]
	1/2	40 [36, 44]	91 [88, 93]
	2/3	76 [72, 80]	65 [61, 69]
	3/4	95 [93, 97]	24 [21, 28]
Set 4-3	0/1	9 [7, 11]	100 [99, 100]
(weighted)	1/2	13 [10, 16]	100 [99, 100]
	2/3	22 [18, 26]	99 [98, 100]
	3/4	32 [28, 36]	97 [95, 98]
	4/5	50 [46, 54]	86 [83, 89]
	5/6	**82 [78, 85]**	**55 [51, 59]**
Set 4-4	0/1	12 [9, 15]	99 [98, 100]
(weighted)	1/2	21 [18, 25]	99 [98, 99]
	2/3	43 [39, 47]	90 [87, 92]
	3/4	77 [73, 80]	65 [61, 69]
	4/5	95 [93, 97]	24 [21, 28]

(Table 5 continued)

	Cut-point	Sensitivity (%) [95% CI]	Specificity (%)[95% CI]
Sets of 3 questions			
Set 3-1	0/1	13 [10, 16]	99 [98, 100]
	1/2	44 [40, 48]	89 [87, 92]
	2/3	**80 [77, 83]**	**56 [52, 60]**
Set 3-2	0/1	9 [7, 12]	100 [99, 100]
	1/2	26 [22, 30]	97 [96, 98]
	2/3	72 [68, 76]	67 [64, 71]
Set 3-3	0/1	12 [9, 15]	100 [99, 100]
	1/2	28 [24, 32]	99 [98, 99]
	2/3	59 [55, 64]	74 [70, 77]
Set 3-4	0/1	13 [10, 16]	99 [98, 100]
(weighted)	1/2	23 [19, 27]	99 [97, 99]
	2/3	45 [40, 49]	89 [86, 91]
	3/4	**80 [77, 83]**	**56 [52, 60]**
Set 3-5	0/1	9 [7, 12]	100 [99, 100]
(weighted)	1/2	13 [10, 16]	100 [99, 100]
	2/3	26 [22, 30]	97 [96, 98]
	3/4	34 [30, 38]	95 [93, 97]
	4/5	72 [68, 76]	67 [64, 71]
Set 3-6	0/1	12 [9, 15]	100 [99, 100]
(weighted)	1/2	13 [10, 16]	100 [99, 100]
	2/3	28 [24, 32]	99 [98, 99]
	3/4	34 [30, 38]	95 [93, 97]
	4/5	59 [55, 64]	74 [70, 77]

(Table 5 continued)

	Cut-point	Sensitivity (%) [95% CI]	Specificity (%)[95% CI]
Sets of 2 questions			
Set 2-1	0/1	16 [13, 19]	99 [98, 99]
	1/2	69 [65, 73]	69 [65, 72]
Set 2-2	0/1	13 [10, 16]	100 [99, 100]
	1/2	34 [30, 38]	95 [93, 97]
Set 2-3	0/1	34 [30, 38]	90 [88, 92]
	1/2	79 [76, 83]	57 [53, 61]
Set 2-4	0/1	16 [13, 19]	99 [98, 99]
(weighted)	1/2	24 [20, 28]	98 [96, 99]
	2/3	69 [65, 73]	69 [65, 72]
Set 2-5	0/1	16 [13, 19]	99 [98, 99]
(weighted)	1/2	24 [20, 28]	98 [96, 99]
	2/3	24 [20, 28]	98 [96, 99]
	3/4	69 [65, 73]	69 [65, 72]

Table 7: External Validation

(1) CSHA-2 data (the primary external validation data base)

	Cut-point	Sensitivity (%) [95% CI]	Specificity (%)[95% CI]
Set 4-1	(0/1)	5 [3, 8]	100 [99, 100]
	(1/2)	16 [13, 21]	99 [97, 100]
	(2/3)	41 [36, 47]	93 [89, 95]
	(3/4)	**80 [75, 84]**	**61 [56, 66]**
Set 3 -1	(0/1)	9 [6, 12]	100 [98, 100]
	(1/2)	35 [30, 40]	93 [91, 96]
	(2/3)	**76 [72, 81]**	**62 [57, 67]**

---

(2) The Ottawa General Hospital Memory Disorder Clinic data

	Cut-point	Sensitivity (%) [95% CI]	Specificity (%)[95% CI]
Set 4-1	(0/1)	21 [17, 26]	99 [95, 100]
	(1/2)	40 [34, 46]	97 [93, 99]
	(2/3)	62 [56, 68]	86 [79, 91]
	(3/4)	**83 [79, 88]**	**60 [51, 68]**
Set 3 -1	(0/1)	25 [20, 30]	98 [94, 99]
	(1/2)	56 [51, 63]	89 [82, 93]
	(2/3)	**81 [76, 86]**	**61 [52, 68]**
MMSE	*(23/24)	59 [53, 65]	89 [82, 93]
	(24/25)	67 [62, 73]	87 [81, 92]
	*(25/26)	75 [71, 81]	82 [75, 88]
	(26/27)	**82 [78, 87]**	**75 [67, 82]**

* (23/24) and (25/26) are the two cut-offs which have been promoted in the published literature and which are employed clinically.

TABLE 9; ROC Analysis and Area Under the Curve (AUC)

Set #	**Non-parametric analysis**		Parametric analysis			
	AUC	*S.E.*	*AUC*	*S.E.*	*Goodness of Fit*	
Sets of 4 questions					Chi Sq.	p
Set 4-1	**.752**	**.014**	.775	.015	3.829	.281
Set 4-2	.759	.014	.776	.014	1.967	.579
Set 4-3	.754	.014	.773	.015	5.551	.352
Set 4-4	.764	.014	.780	.014	5.997	.199
Sets of 3 questions						
Set 3-1	**.742**	**.015**	.769	.015	0.793	.673
Set 3-2	.730	.015	.769	.017	0.672	.714
Set3-3	.698	.016	.696	.022	1.545	.462
Set 3-4	.746	.015	.770	.015	4.59	.204
Set 3-5	.736	.015	.768	.017	1.544	.819
Set 3-6	.699	.016	"Goodness of Fit determines that parametric analysis should not be done"			
Sets of 2 questions						
Set 2-1	.709	.018	.756	.018	0.000	1
Set 2-2	.646	.017	.660	.070	0.000	1
Set 2-3	.716	.015	.752	.016	0.000	1
Set 2-4	.719	.015	.757	.018	1.026	.599
Set 2-5	.719	.015	.757	.018	1.026	.599

Table 10; Responses to the pilot survey of Family Physicians' opinions regarding the format of tools to screen for cognitive impairment (N = 6 Élisabeth Bruyère academic Family Physicians).

Screening Tests; 1 = MMSE, 2 = Clock Drawing (with formal scoring system), 3 = The Ottawa unweighted 3D screen, 4 = The Ottawa weighted 3D screen, 5 = The Ottawa Cognitive Screening Rules - screening algorithm (see survey in Appendix A for examples of the screens)

	SCREENING TEST				
(A) Limitations of this test / barriers to routine use **(please check-off all of options 1 - 6 that apply, if none of options 1-6 apply then select option 7)**	**1**	**2**	**3**	**4**	**5**
□ 1. Too long	2	1			2
□ 2. Requires props (ex. Pencil, paper, drawing...)	3	3			
□ 3. Scoring is too complex to use	1	1		1	
□ 4. Scoring will be too complex to recall	1	4			
□ 5. Will not be acceptable to patients	1				
□ 6. Other limits/barriers (please specify); _____	1	1	1	1	2
OR					
□ 7. No significant limitations or barriers to routine use	2		5	5	2
(B) Please rate the acceptability of this tool in daily clinical practice					
□ 1. Completely acceptable (could routinely use with most if not all patients)	2	1	3	3	1
□ 2. High level of acceptability (could routinely use with many patients)	2	2	3	2	2
□ 3. Equivocal / Neutral		2		1	3
□ 4. Low level of acceptability (possible but unlikely to be used routinely)	2	1			
□ 5. Unacceptable (impossible to use routinely)					
(C) I would be comfortable routinely using this screen for cognitive impairment with my elderly patients.					
□ 1. YES	3	4	6	6	3
□ 2. NO	3	2			3
(D) I already routinely employ this screening tool in my daily practice.(only applies to screens 1 & 2)					
□ 1. YES	4	3			
□ 2. NO	2	3			

Table 11; Responses to the pilot survey of Family Physicians' opinions regarding the format of tools to screen for cognitive impairment (N = 6 Élisabeth Bruyère academic Family Physicians).

Q5. Please rank the 5 screening tools from the Most Likely (#1) to the Least Likely (#5) to be routinely used in your clinical practice.

	Individual Physicians' rankings					
	MD#1,	MD#2,	MD#3,	MD#4,	MD#5,	MD#6
(1) The Folstein Mini-mental Status Examination (MMSE)	2	4	5	1	5	1
(2) The Clock Drawing (using an established method of scoring)	3	5	3	4	4	2
(3) The "Ottawa unweighted 3D Screen"	4	1	1	2	1	5
(4) The "Ottawa weighted 3D Screen"	5	2	2	3	3	4
(5) The "Ottawa Cognitive Screening Rules"	1	3	4	5	2	3

Figure 1 ; Basic Recursive Partitioning strategy #1

(Continue to ask questions if passed previous question - Rightward algorithm).

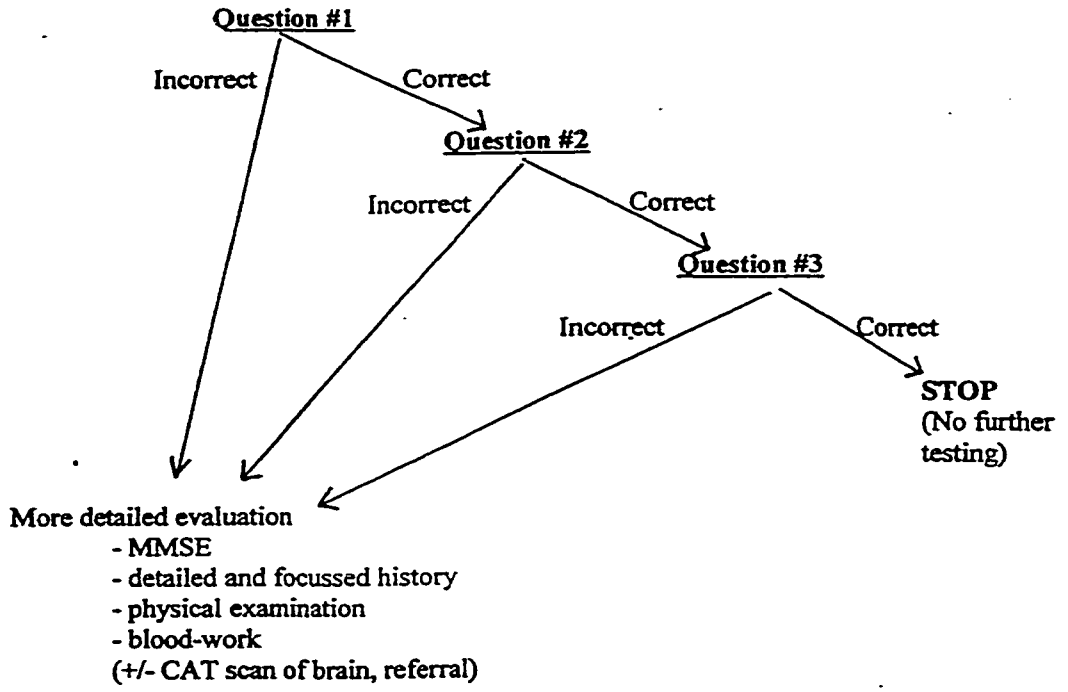


Figure 2; Basic Recursive Partitioning strategy #2

(Continue to ask questions if failed previous question - Leftward algorithm)

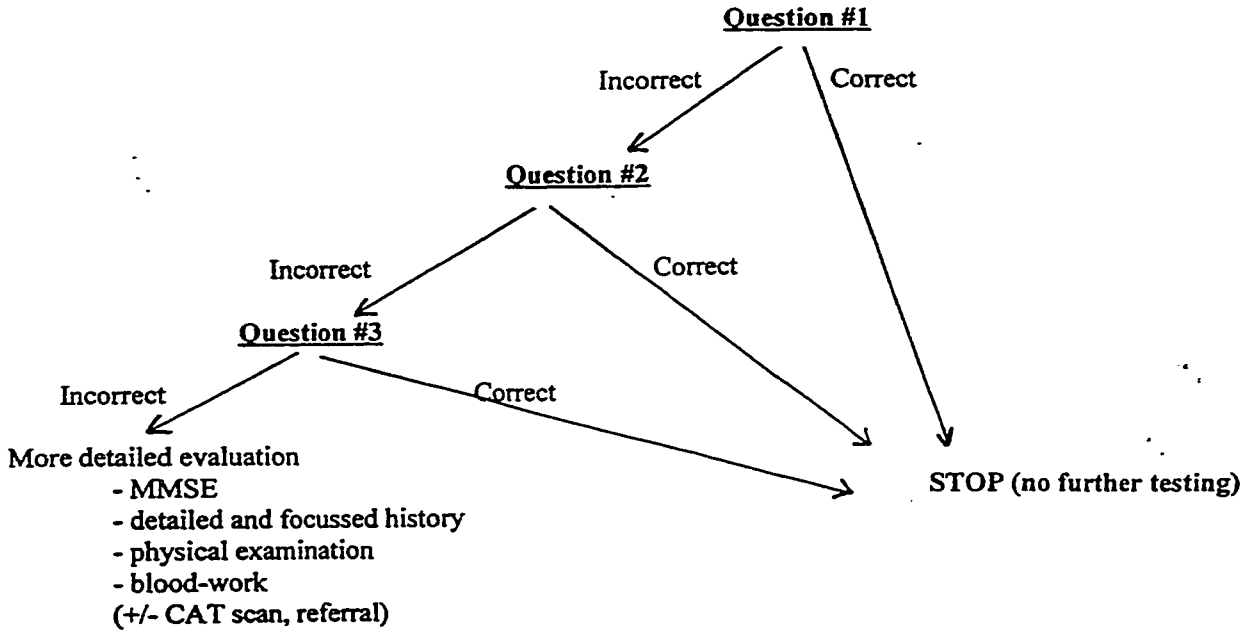


Figure 4: External validation; The CSHA-2 data
Recursive Partitioning Strategy # 1B

1. Always pick the most Sensitive question (Sens max)

(B) Only those who pass (CORRECT answer) the previous question go on to the next question. Those who answer a question incorrectly are considered cognitively impaired and Are automatically required to go for further testing (indicated by 'STOP - further testing').

** This strategy is similar to that employed in the derivation of the Ottawa Ankle Rules

		(Cumulative) Sensitivity (%)	& Specificity (%)
Question #1 ; Spell WORLD backwards.			
INCORRECT (STOP - further testing)	CORRECT	48%	79%
Question #2 ; What is today's date ?			
INCORRECT (STOP - further testing)	CORRECT	74%	64%
Question #3 ; What is this called (point to knuckle) ?			
INCORRECT (STOP - further testing)	CORRECT	**77%**	**60%**

Figure 5: External validation; The CSHA-2 data

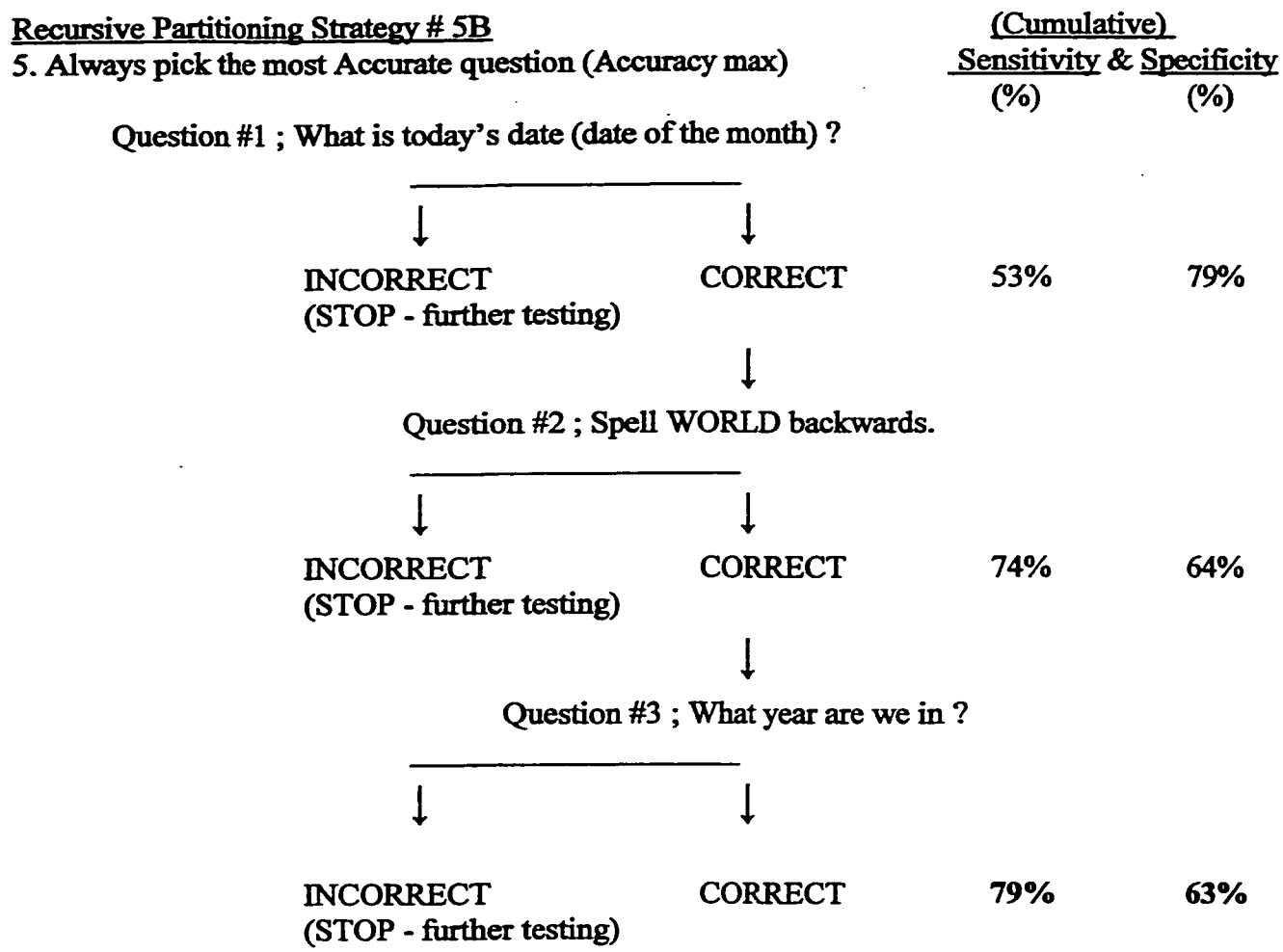


Figure 6; External validation; The Ottawa General Hospital Memory Disorder Clinic data
Due to the fact that strategy 1B (and hence 4B) contained a 3MS question not found in the MMSE, it is not possible to validate recursive partitioning strategies 1B and 4B in this data base.

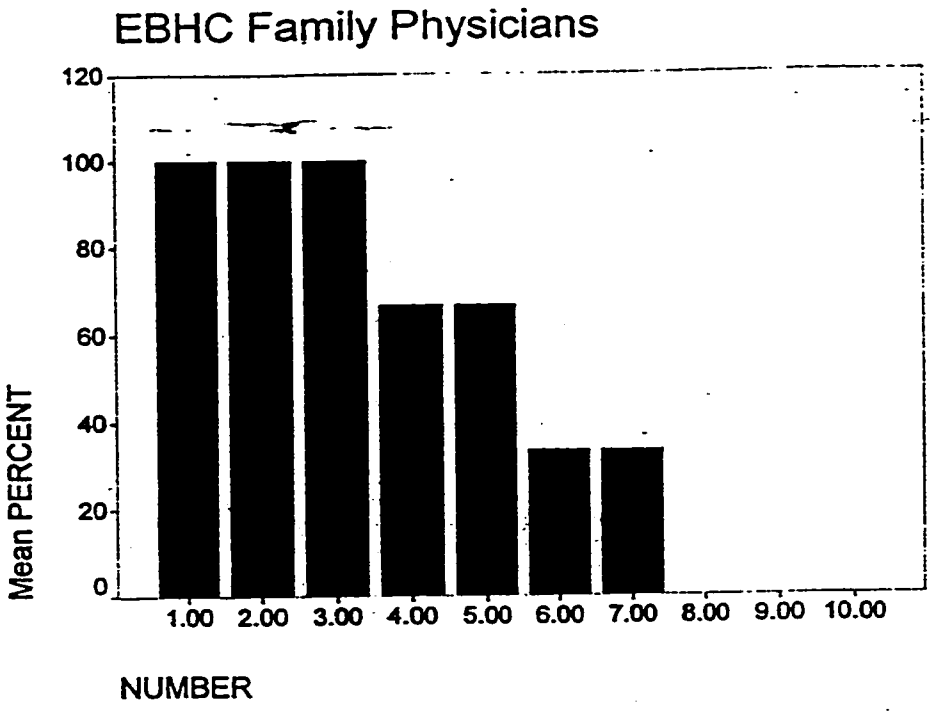
Recursive Partitioning Strategy # 5B

(Cumulative)

5. Always pick the most Accurate question (Accuracy max)

		Sensitivity (%)	Specificity (%)
Question #1 ; What is today's date (date of the month) ?			
INCORRECT (STOP - further testing)	CORRECT	66%	70%
Question #2 ; Spell WORLD backwards.			
INCORRECT (STOP - further testing)	CORRECT	80%	63%
Question #3 ; What year are we in ?			
INCORRECT (STOP - further testing)	CORRECT	**81%**	**62%**

Figure 7 The maximum number of cognitive questions which physicians surveyed felt they would be able to recall



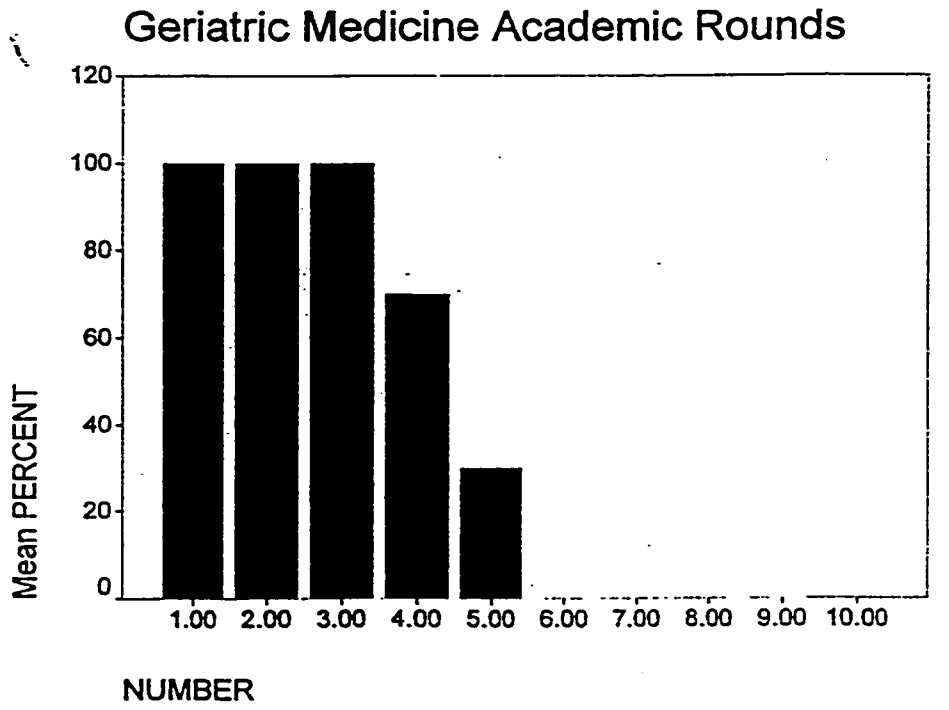


Figure 8 : The Sensitivity and Specificity of different MMSE Cut-off Scores
(Derived from the Ottawa General Hospital Memory Disorder Clinic)

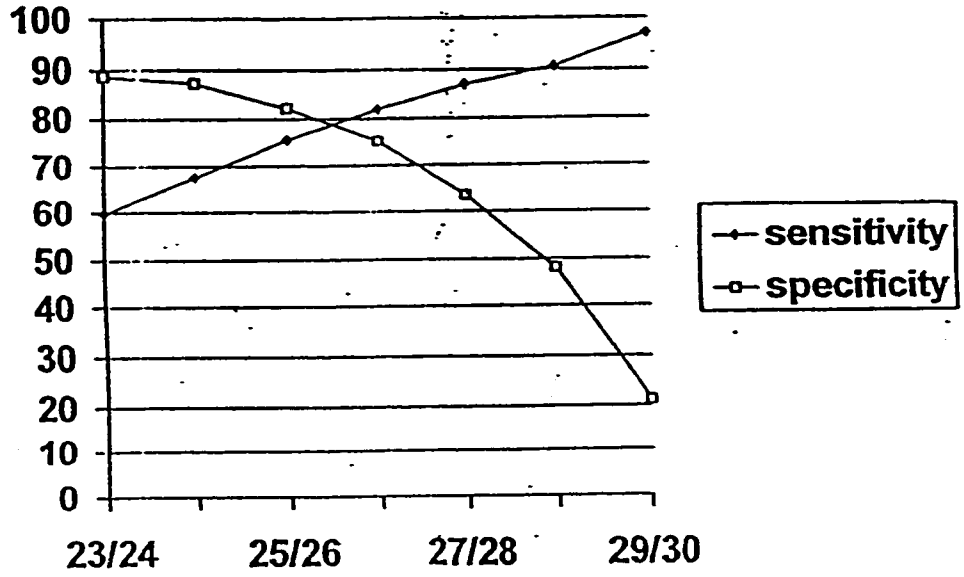
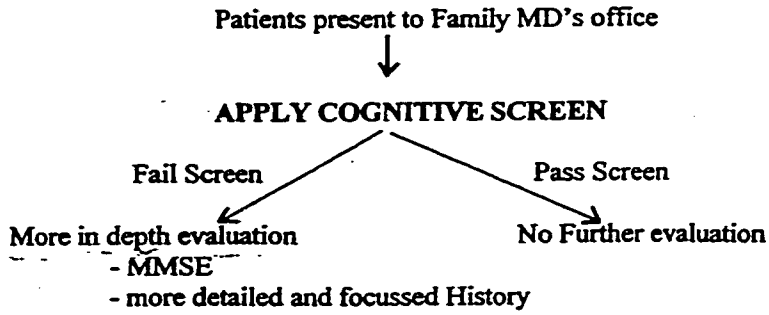
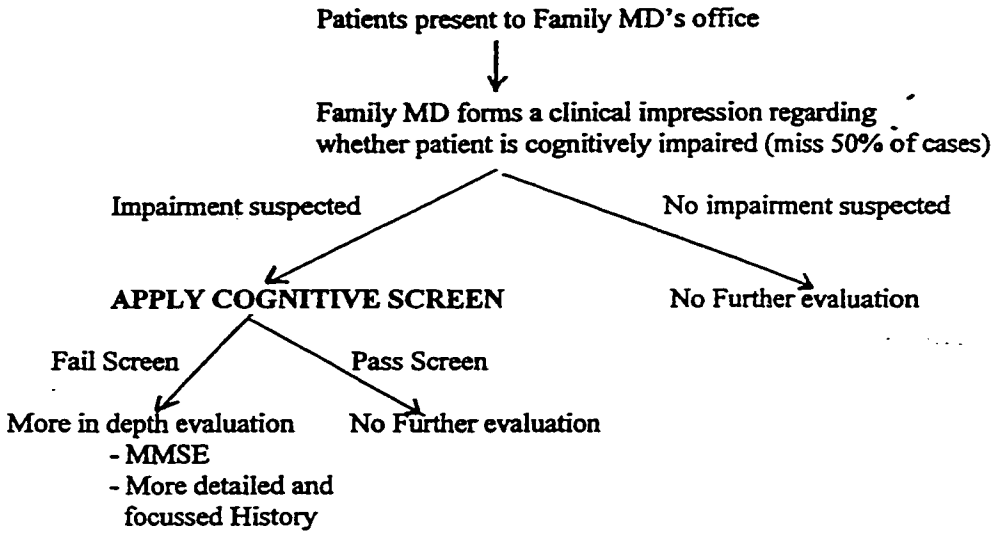


Figure 9: The possible roles of screening for cognitive impairment.

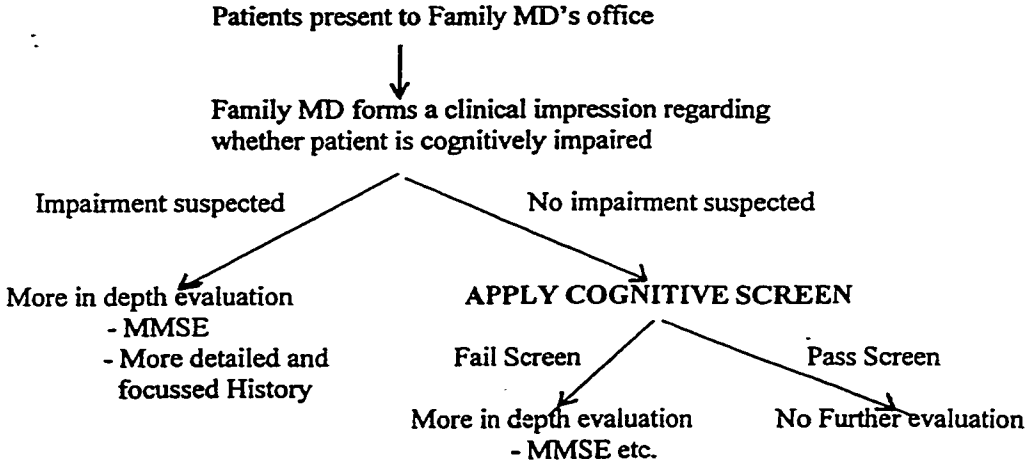
Scenario #1: SCREEN ALL PATIENTS (Parallels the application of screen in this study)



Scenario #2 : ONLY THOSE PATIENTS SUSPECTED OF COGNITIVE IMPAIRMENT ARE SCREENED (suggested by some MDs)



Scenario #3 : ONLY THOSE PATIENTS IN WHOM COGNITIVE IMPAIRMENT IS NOT SUSPECTED ARE SCREENED (The recommended use of screening)



APPENDICES

Appendix A

Comparison of the Quality of Mental Status Tests*

Measurements	Scale	Number of items	Application	Administered by (time)	Studies using method	Reliability		Validity	
						Thoroughness	Results	Thoroughness	Results
Dementia Rating Scale (Mattis, 1973)	ordinal	22	research	expert	several	++	++	++	++
Cognitive Capacity Screening Examination (Jacobs, 1977)	ordinal	30	clinical	expert (5–15 min)	several	+	++	++	++
Clock Drawing Test (Various authors, 1986)	ordinal	1†	clinical, screening	self	many	++	++	++	++
Alzheimer's Disease Assessment Scale (Rosen, 1983)	ordinal	21	screening	expert	several	+	+++	++	++
Information-Memory-Concentration Test (Blessed, 1968)	ordinal	29	clinical, survey	staff (10 min)	many	+	+++	++	+++
Dementia Scale (Blessed, 1968)	ordinal	22	research	interviewer	several	+	+	++	+
Mental Status Questionnaire (Kahn, 1960)	ordinal	10	clinical, survey	interviewer (5–10 min)	several	+	++	++	++
Short Portable Mental Status Questionnaire (Pfeiffer, 1975)	ordinal	10	screening, survey	interviewer (2 min)	many	+	++	++	++
Mini-Mental Status Examination (Folstein, 1975)	ordinal	30	clinical, screening	interviewer (5–10 min)	many	+++	++	+++	+++
Clifton Assessment Procedures for the Elderly (Pattie, 1975)	ordinal	12	clinical	staff	several	++	++	++	++
Cambridge Mental Disorders of the Elderly Examination (CAMDEX) (Roth, 1986)	ordinal	57	research, clinical	expert (80 mins)	many	+	+++	++	+++

Appendix B

Watson et al. 1993

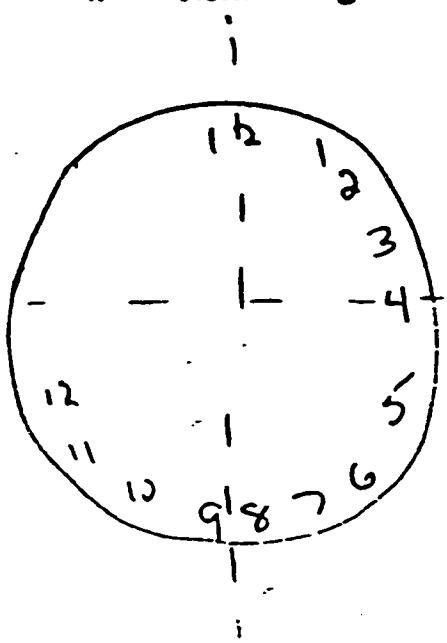
The patient is instructed to draw the numbers within a predrawn circle to make the circle look like the face of a clock

1. Divide the circle into 4 equal quadrants by drawing one line through the center of the circle and the number 12(or mark that best corresponds to the 12) and a second line perpendicular to and bisecting the first

2. Count the number of digits in each quadrant in the clockwise direction beginning with the digit corresponding to number 12. Each digit is counted only once. If a digit falls on one of the reference lines, it is included in the quadrant that is clockwise to the line. Any three digits in a quadrant is considered to be correct.

3. For and error in the number of digits in the first, second, or third quadrants assign a score of 1. For any error in the number of digits in the fourth quadrant assign a score of 4.

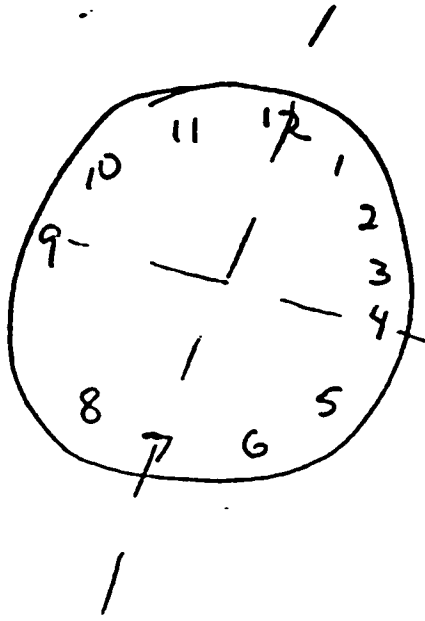
4. Normal range of score is 0-3. Abnormal(demented) score is 4-7.



Errors.

Q_1 1
Q_2 1
Q_3 1
Q_4 4

total 7



Q_1 1
Q_2 0
Q_3 1
Q_4 0

total 2.

Appendix C

Administration and scoring:
The Sunderland method

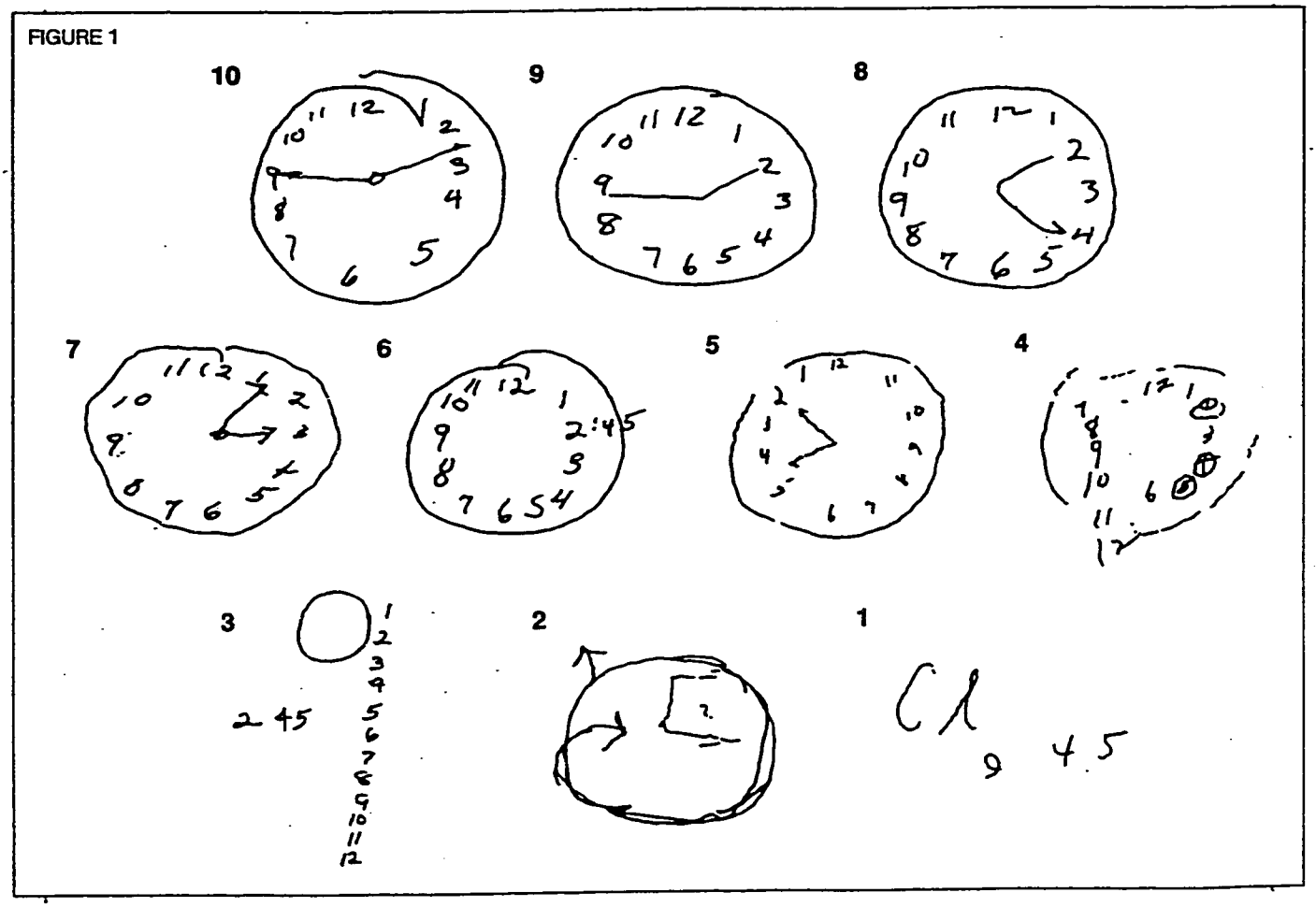
1. Ask the person to draw a clock with all the numbers on it.
2. Ask the person to draw hands on the face of the clock to make it read 2:45.
3. Repeat the instructions in exactly the same wording as often as necessary, but do not give any other directions. Do not cover up any clocks or other timepieces in the room.
4. Score the drawing according to the criteria shown in Table 1 and Figure 1. Generally, a score of 5 or less is consistent with a diagnosis of AD.

10. Hands are in correct position (the hour hand approaches 3 o'clock).
9. Slight errors appear in placement of the hands.
8. More noticeable errors appear in placement of the hands.
7. Placement of the hands is significantly off course.
6. The hands are used inappropriately (digital display or circled numbers).

Circle and numbers are generally not intact

5. Numbers are crowded at one end of the clock or are reversed. Hands may still appear in some fashion.
4. The number sequence is further distorted. The integrity of the clock face is now gone (numbers are missing or placed outside the clock face).
3. The numbers and clock face are no longer connected in the drawing. The hands have been omitted.
2. The drawing shows some evidence of instructions having been received, but the representation of a clock is vague.
1. There is no attempt at all or only an uninterpretable effort at a drawing.

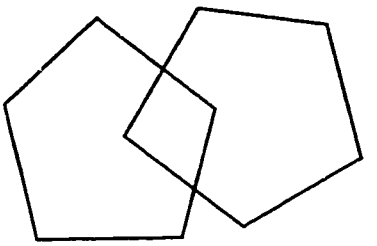
FIGURE 1



Appendix D

Mini-Mental State Examination (MMSE)[1,2*]

Make the patient comfortable and establish rapport. Ask questions in the order listed.
Total possible score is 30.

Maximum Score	Score		
		ORIENTATION	
5	( )	What is the (year) (season) (date) (day) (month)?	
5	( )	Where are we (state) (county) (town or city) (hospital) (floor)?	
		REGISTRATION	
3	( )	Name 3 common objects (eg, "apple," "table," "penny"). Take 1 second to say each. Then ask the patient to repeat all 3. Give 1 point for each correct answer. Then repeat them until he/she learns all 3. Count trials and record. Trials: _____	
		ATTENTION AND CALCULATION	
5	( )	Serial 7's backwards. Stop after 5 answers. Alternatively, spell "WORLD" backwards. The score is the number of letters in correct order (D____L____R____O____W____).	
		RECALL	
3	( )	Ask for the 3 common objects named during registration above. Give 1 point for each correct answer. [Note: recall cannot be tested if all 3 objects were not remembered during registration.]	
		LANGUAGE	
2	( )	Name a "pencil" and "watch."	(2 points)
1	( )	Repeat the following: "No ifs, ands, or buts."	(1 point)
3	( )	Follow a 3-stage command: "Take a paper in your right hand, fold it in half, and put it on the floor."	(3 points)
1	( )	Read and obey the following: CLOSE YOUR EYES.	(1 point)
1	( )	Write a sentence.	(1 point)
1	( )	Copy the following design: 	(1 point)

Maximum Total Score	Total Score	*Suggested guideline[3] for determining the severity of cognitive impairment:*
30	_____	Mild: MMSE ≥21
		Moderate: MMSE 10-20
		Severe: MMSE ≤9

Expected decline in MMSE scores in untreated mild to moderate Alzheimer's patient is 2 to 4 points per year.[4]

– (3MS) PROFESSION: _____

Edu _____ (Years)

3MS

___/5 DATE & PLACE OF BIRTH

Date yr ____, mo ____, D ____.
Place Town _____ Prov.____

___/3 REGISTRATION
(No. of presentations) _____

SHIRT-BROWN-HONESTY
(or SHOES-BLACK-MODESTY)
(or SOCKS-BLUE-CHARITY)

___/7 MENTAL REVERSAL
2 Accurate 5 to 1
1 – 1 or 2 errors/misses
D L R O W (WORLD)
__ __ __ __ __

___/9 FIRST RECALL

3 Spontaneous recall
2 After "something to wear"
1 "SHOES-SHIRT-SOCKS"
3 Spontaneous recall
2 After "a colour"
1 "BLUE-BLACK-BROWN"

3 Spontaneous recall
2 After "a good personal quality"
1 "HONESTY-CHARITY-MODESTY"

___/15 TEMPORAL ORIENTATION

Year
8 Accurate
4 Missed by 1 year
2 Missed by 2-5 years
Season
1 Accurate or within 1 month
Month
2 Accurate or within 5 days
1 Missed by 1 month
Day of Month
3 Accurate
2 Missed by 1 or 2 days
1 Missed by 3-5 days
Day of Week
1 Accurate

___/5 SPATIAL ORIENTATION
2 Home Address
1 General
1 City (town)
1 Hospital

___/5 NAMING (MMS: Pencil ___ Watch ___)
Forehead ___, Chin ___, Shoulder
Elbow ___, Knuckle ___.

___/10 FOUR-LEGGED ANIMALS (30 SEC)
1 point each

___/6 SIMILARITIES

Arm–Leg
2 Body part; limb etc
1 Less correct answer
Laughing–Crying
2 Feeling; emotion
1 Other correct answer
Eating–Sleeping
2 Essential for life
1 Other correct answer

___/5 REPETITION

(2) "I WOULD LIKE TO GO HOME/OUT"
(1) 1 or 2 missed/wrong words
"NO IFS ___ ANDS ___ BUTS ___"

___/3 READ & OBEY "CLOSE YOUR EYES"
3 Obeys without prompting
2 Obeys after prompting
1 Reads aloud only
(spontaneously or by request)

___/5 WRITING (1 minute)

___/10 COPYING TWO PENTAGONS (1 min)

4/4 5 approx. equal sides
3/3 5 unequal (>2:1) sides
2/2 Other enclosed figure
1/1 2 or more lines
2 4 corners
1 Not-4-corner enclosure

___/3 THREE-STAGE COMMAND

___ TAKE THIS PAPER WITH YOUR LEFT/RIGHT HAND
___ FOLD IN HALF, AND
___ HAND IT BACK TO ME

___/9 SECOND RECALL
(Something to wear)
(Colour)
(Good personal quality)

APPENDIX F

The CAGE questionnaire

	POINTS
1. Have you ever *C*ut down on your drinking ?	1
2. Do people ever *A*nnoy you about your drinking ?	1
3. Do you ever feel *G*uilty about your drinking ?	1
4. Do you ever need an *E*ye-opener ?	1
TOTAL =	__ / 4

APPENDIX G

The Goldman Cardiac Risk Index

	POINTS
History	
Myocardial Infarction within the last 6 months	10
Age over 70	5
Physical Examination	
S3 or Jugular Venous distention	11
Important aortic stenosis	3
ECG	
Rhythm other than sinus or sinus + APBs	7
More than 5 PVCs per minute	7
Poor general medical status	3
- PO2<60 or PCO2>50 mmHg K<3 or HCO3<20, BUN>50 or Cr>3 mg/dl Abnormal AST, chronic liver disease, bedridden	
Operational	
Intraperitoneal, intrathoracic or aortic operation	3
Emergency operation	4
TOTAL POSSIBLE	53

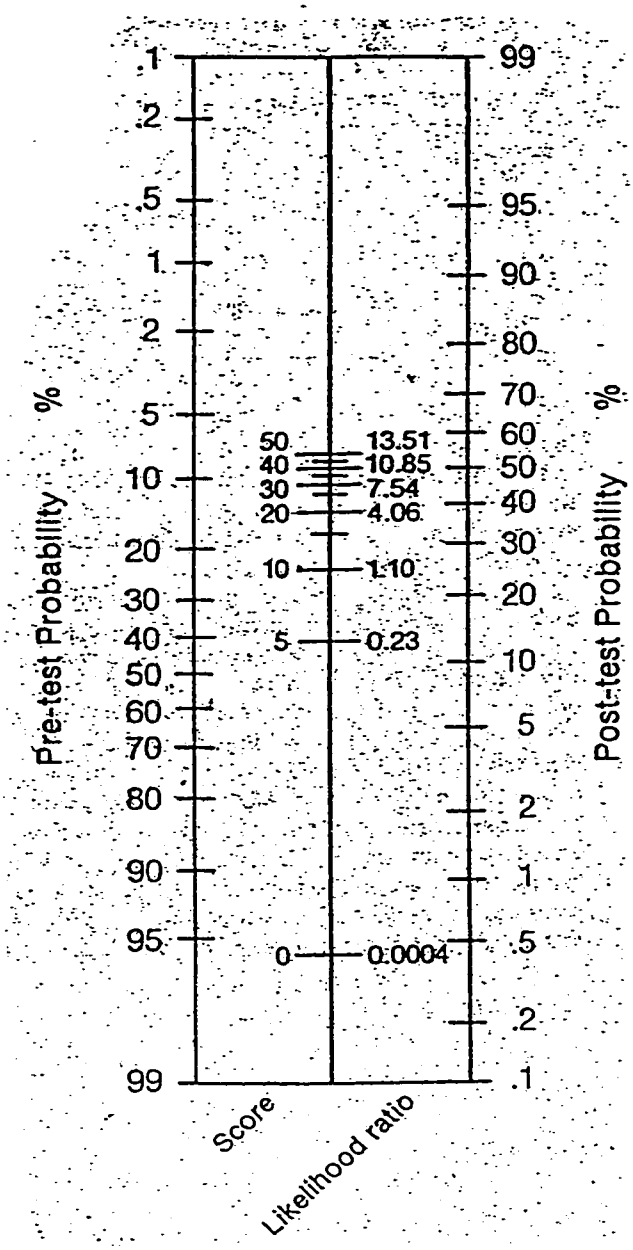
APPENDIX H - Detsky's Modified Cardiac index

Obtain modified cardiac score from scale on the left. Then anchor a straight edge on the pretest probability line of the likelihood ratio nomogram (left line). Direct the straight edge through the point in the center line reflecting the patient's modified cardiac score. Where the edge meets the right-handed column one may determine the patient's risk of peri-operative cardiac complications.

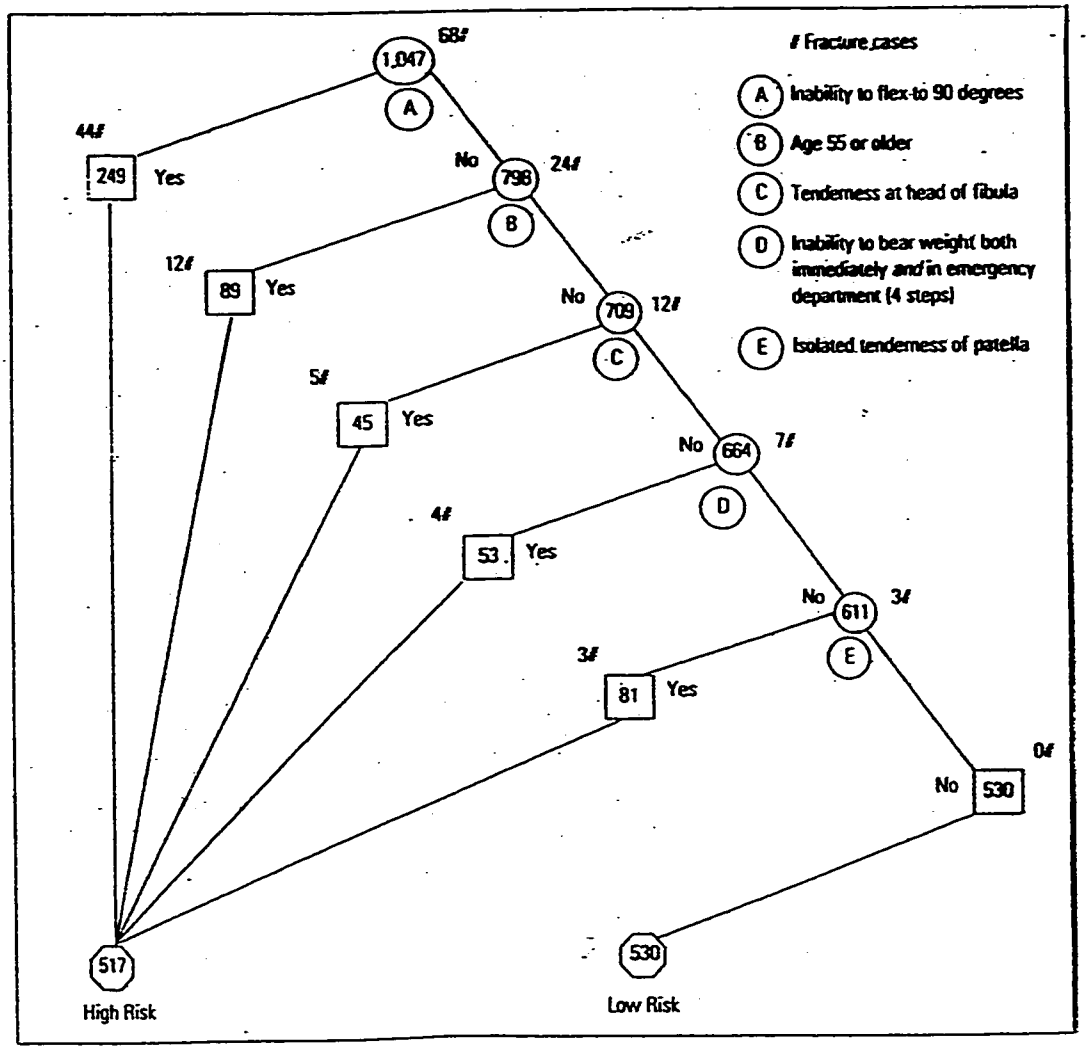
Modified Multifactorial Index

	Points
Coronary artery disease	
Myocardial infarction within 6 months	10
Myocardial infarction more than 6 months	5
Canadian Cardiovascular Society angina	
Class III	10
Class IV	20
Unstable angina within 6 months	10
Alveolar pulmonary edema	
Within 1 week	10
Ever	5
Valvular disease	
Suspected critical aortic stenosis	20
Arrhythmias	
Rhythm other than sinus or sinus plus APBs* on last preoperative electrocardiogram	5
More than five premature ventricular contractions at any time prior to surgery	5
Poor general medical status†	5
Age over 70	5
Emergency operation	10

* APB = atrial premature beat.
† As defined in original multifactorial index (Table 1).



APPENDIX I - Example of χ^2 recursive partitioning analysis to derive a decision rule (the Ottawa Knee Rules)



APPENDIX J

Properties derived from a 2 X 2 contingency table

		Gold Standard Result	
		Disease present	Disease absent
Test Result	Positive	a	b
	Negative	c	d

[1] Sensitivity = a / (a + c) ; proportion of people with disease who have a positive test

[2] Specificity = d / (b + d) ; proportion of people free of disease who have a negative test

[3] Positive Predictive Value (PPV) = a / (a+b) ; proportion of people with a positive test who have the disease

[4] Negative Predictive Value (NPV) = d/(c+d) ; proportion of people with a negative test who do not have the disease

[5] Accuracy = (a+d) / (a + b + c + d) ; proportion of people properly categorized by test
These basic features were employed in the development of the most basic and reproducible recursive partitioning strategies.

APPENDIX K

PILOT SURVEY OF
FAMILY PHYSICIANS' OPINIONS REGARDING THE FORMAT OF TOOLS TO SCREEN FOR COGNITIVE IMPAIRMENT

Dr. Frank Molnar
Medical Director EBHC Geriatric Day Hospital
Medical Director Geriatric Outreach Team

1 / May / 2000

Dr. _____

I would greatly appreciate it if you could complete the following survey. The length, format and survey itself have been modified and approved by Drs. Bernstein and Elmslie. It is my hope that this survey will lead to future collaborative studies into screening in primary care.

I would value your opinions regarding the likelihood that you would use the following clinical tools to screen for cognitive impairment in your clinical practice. Copies of the respective screening tools are attached. The first 2 tools (MMSE, Clock Drawing) are already available for clinical practice. The last 3 tools (Ottawa 3D unweighted screen, Ottawa 3D weighted screen, Ottawa Cognitive Screening Rules) have been developed and demonstrate sensitivities and specificities similar to those of the first 2 tools.

This survey is designed to focus on the acceptability of these screens to busy clinicians and the probability of routine use. Please review all 5 screening tools before answering the questions.

Completed surveys should be forwarded to me at the EBHC Geriatric Day Hospital via internal mail. I would like to thank you in advance for your assistance.

Frank Molnar

1. Approximately what percentage of your office-based community practice patients are seniors (65 years or older) ?

_____ %

2. Approximately how many seniors (patients 65 years or older) do you see per week in your office-based practice ?

_____ (# of patients)

3. ESTABLISHED SCREENS FOR COGNITIVE IMPAIRMENT

(1) **The Folstein Mini-mental Status Examination (MMSE)** - SAMPLE ON P.8

(A) Limitations of this test / barriers to routine use **(please check-off all of options 1 - 6 that apply, if none of options 1-6 apply then select option 7)**

- ☐ 1. Too long
- ☐ 2. Requires props (ex. Pencil, paper, drawing...)
- ☐ 3. Scoring is too complex to use
- ☐ 4. Scoring will be too complex to recall
- ☐ 5. Will not be acceptable to patients
- ☐ 6. Other limits/barriers (please specify); _____

OR

- ☐ 7. No significant limitations or barriers to routine use

(B) Please rate the acceptability of this tool in daily clinical practice

- ☐ 1. Completely acceptable (could routinely use with most if not all patients)
- ☐ 2. High level of acceptability (could routinely use with many patients)
- ☐ 3. Equivocal / Neutral
- ☐ 4. Low level of acceptability (possible but unlikely to be used routinely)
- ☐ 5. Unacceptable (impossible to use routinely)

(C) I would be comfortable routinely using this screen for cognitive impairment with my elderly patients.

- ☐ 1. YES
- ☐ 2. NO

(D) I already routinely employ this screening tool in my daily practice.

- ☐ 1. YES
- ☐ 2. NO

(2) **The Clock Drawing** (using an established method of scoring)

*** SAMPLE OF CLOCK DRAWING SCORING METHOD on P.9***

(A) Limitations of this test / barriers to routine use **(please check-off all of options 1 - 6 that apply, if none of options 1-6 apply then select option 7)**

- ☐ 1. Too long
- ☐ 2. Requires props (ex. Pencil, paper, drawing...)
- ☐ 3. Scoring is too complex to use
- ☐ 4. Scoring will be too complex to recall
- ☐ 5. Will not be acceptable to patients
- ☐ 6. Other limits/barriers (please specify); _____

OR

- ☐ 7. No significant limitations or barriers to routine use

(B) Please rate the acceptability of this tool in daily clinical practice

- ☐ 1. Completely acceptable (could routinely use with most if not all patients)
- ☐ 2. High level of acceptability (could routinely use with many patients)
- ☐ 3. Equivocal / Neutral
- ☐ 4. Low level of acceptability (possible but unlikely to be used routinely)
- ☐ 5. Unacceptable (impossible to use routinely)

(C) I would be comfortable routinely using this screen for cognitive impairment with my elderly patients.

- ☐ 1. YES
- ☐ 2. NO

(D) I already routinely employ this screening tool in my daily practice.

- ☐ 1. YES
- ☐ 2. NO

4. NEW SCREENING TOOLS FOR COGNITIVE IMPAIRMENT

(3) **The "Ottawa unweighted 3D Screen"**

*****SAMPLE*****

	POINTS
D ay (What is the day of the week)	**1**
D ate (What is today's date - ex " the 21st of the month")	**1**
D lrow (spell WORLD backwards)	**1**

Scoring - anyone scoring less than a perfect score fails the test and goes on to more detailed history and cognitive testing

(A) Limitations of this test / barriers to routine use **(please check-off all of options 1 - 6 that apply, if none of options 1-6 apply then select option 7)**

- ☐ 1. Too long
- ☐ 2. Requires props (ex. Pencil, paper, drawing...)
- ☐ 3. Scoring is too complex to use
- ☐ 4. Scoring will be too complex to recall
- ☐ 5. Will not be acceptable to patients
- ☐ 6. Other limits/barriers (please specify); _____

OR

- ☐ 7. No significant limitations or barriers to routine use

(B) Please rate the acceptability of this tool in daily clinical practice

- ☐ 1. Completely acceptable (could routinely use with most if not all patients)
- ☐ 2. High level of acceptability (could routinely use with many patients)
- ☐ 3. Equivocal / Neutral
- ☐ 4. Low level of acceptability (possible but unlikely to be used routinely)
- ☐ 5. Unacceptable (impossible to use routinely)

(C) I would be comfortable routinely using this screen for cognitive impairment with my elderly patients.

- ☐ 1. YES
- ☐ 2. NO

(4) The "Ottawa weighted 3D Screen"

*****SAMPLE*****

	POINTS
D ay (What is the day of the week)	**2**
D ate (What is today's date - ex " the 1st of the month")	**1**
D lrow (spell WORLD backwards)	**1**

Scoring - anyone scoring less than 3 fails the test and goes on to more detailed history and cognitive testing

(A) Limitations of this test / barriers to routine use **(please check-off all of options 1 - 6 that apply, if none of options 1-6 apply then select option 7)**

- ☐ 1. Too long
- ☐ 2. Requires props (ex. Pencil, paper, drawing...)
- ☐ 3. Scoring is too complex to use
- ☐ 4. Scoring will be too complex to recall
- ☐ 5. Will not be acceptable to patients
- ☐ 6. Other limits/barriers (please specify); _____

OR

- ☐ 7. No significant limitations or barriers to routine use

(B) Please rate the acceptability of this tool in daily clinical practice

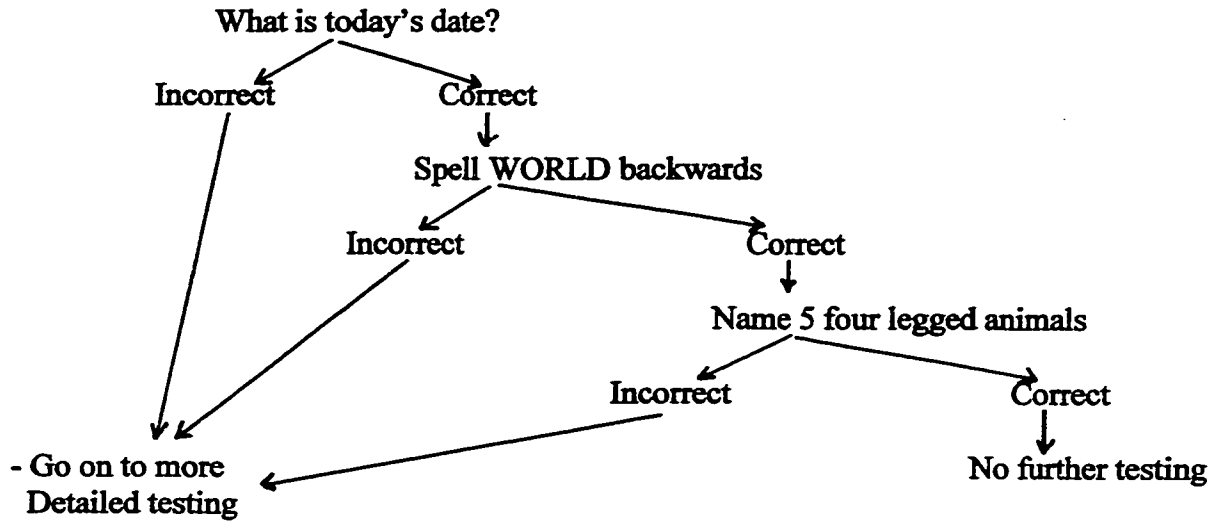
- ☐ 1. Completely acceptable (could routinely use with most if not all patients)
- ☐ 2. High level of acceptability (could routinely use with many patients)
- ☐ 3. Equivocal / Neutral
- ☐ 4. Low level of acceptability (possible but unlikely to be used routinely)
- ☐ 5. Unacceptable (impossible to use routinely)

(C) I would be comfortable routinely using this screen for cognitive impairment with my elderly patients.

- ☐ 1. YES
- ☐ 2. NO

(5) **The "Ottawa Cognitive Screening Rules"**

SAMPLE



(A) Limitations of this test / barriers to routine use **(please check-off all of options 1 - 6 that apply, if none of options 1-6 apply then select option 7)**

- ☐ 1. Too long
- ☐ 2. Requires props (ex. Pencil, paper, drawing...)
- ☐ 3. Scoring is too complex to use
- ☐ 4. Scoring will be too complex to recall
- ☐ 5. Will not be acceptable to patients
- ☐ 6. Other limits/barriers (please specify); _____

OR

- ☐ 7. No significant limitations or barriers to routine use

(B) Please rate the acceptability of this tool in daily clinical practice

- ☐ 1. Completely acceptable (could routinely use with most if not all patients)
- ☐ 2. High level of acceptability (could routinely use with many patients)
- ☐ 3. Equivocal / Neutral
- ☐ 4. Low level of acceptability (possible but unlikely to be used routinely)
- ☐ 5. Unacceptable (impossible to use routinely)

(C) I would be comfortable routinely using this screen for cognitive impairment with my elderly patients.

- ☐ 1. YES
- ☐ 2. NO

5. Please rank the 5 screening tools from the Most Likely (#1) to the Least Likely (#5) to be routinely used in your clinical practice.

Most to least likely

(1) The Folstein Mini-mental Status Examination (MMSE) # _____

(2) The Clock Drawing (using an established method of scoring) # _____

(3) The "Ottawa unweighted 3D Screen" # _____

(4) The "Ottawa weighted 3D Screen" # _____

(5) The "Ottawa Cognitive Screening Rules" # _____

6. Please indicate the maximum number of cognitive questions you would be able to recall and apply in clinical practice (without the use of an acronym such as ABCs, CAGE or APGAR). Only check off a **single** option.

- ☐ 1. Maximum of one cognitive question
- ☐ 2. Maximum of two cognitive questions
- ☐ 3. Maximum of three cognitive questions
- ☐ 4. Maximum of four cognitive questions
- ☐ 5. Maximum of five cognitive questions
- ☐ 6. Maximum of six cognitive questions
- ☐ 7. Maximum of seven cognitive questions
- ☐ 8. Maximum of eight cognitive questions
- ☐ 9. Maximum of nine cognitive questions
- ☐ 10. Maximum of ten cognitive questions

Thank you very much for your assistance. If you have any other opinions / ideas please feel free to list them below. It is my hope that some of these ideas may lead to future collaborative studies into the development of practical and useful screening tools for primary care. If you have ideas regarding such studies please list them below.

_____

_____

_____

_____

Frank Molnar

APPENDIX L : Results of the 10 Basic Recursive Partitioning Analytic Schemes

Recursive Partitioning Strategy # 1A

1. Always pick the most Sensitive question (Sens max)
   (A) Only those who fail (INCORRECT answer) the previous question go on to the next question. Those who answer a question correctly are considered cognitively normal and do not require further testing (indicated by 'STOP').

		Sensitivity (%) [95% CI]	Specificity
Question #1 ; Spell WORLD backwards.			
↓ INCORRECT	↓ CORRECT (STOP)	61 [56, 65]	70 [66, 73]
Question #2 ; What is today's date?			
↓ INCORRECT	↓ CORRECT (STOP)	34 [30, 38]	90 [88, 92]
Question #3 ; What year are we presently in?			
↓ INCORRECT	↓ CORRECT (STOP)	13 [10, 16]	100 [99, 100]

Question #4 ; No further statistically significant splits (i.e. questions)

↓

Go on for more detailed
Cognitive evaluation

APPENDIX L (cont.) ; Recursive Partitioning Strategy # 1B

1. Always pick the most Sensitive question (Sens max)

(B) Only those who pass (CORRECT answer) the previous question go on to the next question. Those who answer a question incorrectly are considered cognitively impaired and are automatically required to go for further testing (indicated by 'STOP- further testing').

Question #1 ; Spell WORLD backwards.		**Sensitivity (%) [95% CI]**	**Specificity**
INCORRECT (STOP - further testing)	CORRECT	61 [56, 65]	70 [66, 73]
Question #2 ; What is today's date ?			
INCORRECT (STOP - Further testing)	CORRECT	79 [76, 83]	57 [53, 61]
Question #3 ; What is this called (point to knuckle) ?			
INCORRECT (STOP - Further testing)	CORRECT	**84 [81, 87]**	**50 [46, 54]**
Question #4 ; Three word recall (requires prior registration)			
INCORRECT (STOP - further testing)	CORRECT	98 [97, 99]	13 [11, 16]

APPENDIX L (cont.) ; Recursive Partitioning Strategy # 2A

2. Always pick the most Specific question (Spec max)

(A) Only those who fail (INCORRECT answer) the previous question go on to the next question. Those who answer a question correctly are considered cognitively normal and do not require further testing (indicated by 'STOP').

Question #1 ; Identify this body part (point to chin).		**Sensitivity (%) [95% CI]**	**Specificity**
↓	↓		
INCORRECT	CORRECT (STOP)	3 [2, 4]	100 [99, 100]
↓			
Question #2 ; What is today's date ?			
↓	↓		
INCORRECT	CORRECT (STOP)	2 [1, 3]	100 [99, 100]
↓			
Question #3 ; No further statistically significant splits (i.e. questions).			
↓			
Go on for more detailed cognitive evaluation			

APPENDIX L (cont.) ; Recursive Partitioning Strategy # 2B

2. Always pick the most Specific question (Spec max)

(B) Only those who pass (CORRECT answer) the previous question go on to the next question. Those who answer a question incorrectly are considered cognitively impaired and are automatically required to go for further testing (indicated by 'STOP - further testing').

Question #1 ; Identify this body part (point to chin).		**Sensitivity (%) [95% CI]**	**Specificity**
INCORRECT (STOP - further testing)	CORRECT	3 [2, 4]	100 [99, 100]
Question #2 ; What month were you born in?			
INCORRECT (STOP - further testing)	CORRECT	6 [4, 8]	99 [98, 100]
Question #3 ; What province are we in ?			
INCORRECT (STOP - further testing)	CORRECT	11 [8, 14]	98 [97, 99]
Question #4 ; What city are we in ?			
INCORRECT (STOP - further testing)	CORRECT	15 [12, 18]	97 [96, 98]

APPENDIX L (cont.) ; Recursive Partitioning Strategy # 3A

3. Always pick the question with the highest PPV (PPV max)

(A) Only those who fail (INCORRECT answer) the previous question go on to the next question. Those who answer a question correctly are considered cognitively normal and do not require further testing (indicated by 'STOP').

		Sensitivity (%) [95% CI]	Specificity
Question #1 ; What month is it now ?			
↓	↓		
INCORRECT	CORRECT (STOP)	15 [12, 18]	99 [97, 99]
↓			
Question #2 ; What year where you born in ?			
↓	↓		
INCORRECT	CORRECT (STOP)	3 [2, 5]	100 [99, 100]
↓			
Question #3 ; No further statistically significant splits (i.e. questions).			
↓			
Go on for more detailed cognitive evaluation			

APPENDIX L (cont.) ; Recursive Partitioning Strategy # 3B

3. Always pick the question with the highest PPV (PPV max)

(B) Only those who pass (CORRECT answer) the previous question go on to the next question. Those who answer a question incorrectly are considered cognitively impaired and are automatically required to go for further testing (indicated by 'STOP - further testing').

		Sensitivity (%) [95% CI]	Specificity
Question #1 ; What month is it now ?			
↓ INCORRECT (STOP - further testing)	↓ CORRECT	15 [12, 18]	99 [97, 99]
↓ Question #2 ; What day of the week is it ?			
↓ INCORRECT (STOP - further testing)	↓ CORRECT	29 [26, 33]	96 [95, 98]
↓ Question #3 ; What province are we in ?			
↓ INCORRECT (STOP - further testing)	↓ CORRECT	32 [28, 36]	96 [94, 97]
↓ Question #4 ; Identify this body part (point to chin).			
↓ INCORRECT (STOP - further testing)	↓ CORRECT	34 [30, 38]	95 [93, 97]

APPENDIX L (cont.) ; Recursive Partitioning Strategy # 4A

4. Always pick the question with the highest NPV (NPV max)

(A) Only those who fail (INCORRECT answer) the previous question go on to the next question. Those who answer a question correctly are considered cognitively normal and do not require further testing (indicated by 'STOP').

** This algorithm is the same as Strategy # 1A - Sensitivity Max

		Sensitivity (%) [95% CI]	Specificity
Question #1 ; Spell WORLD backwards.			
↓	↓		
INCORRECT	CORRECT (STOP)	61 [56, 65]	70 [66, 73]
↓			
Question #2 ; What is today's date ?			
↓	↓		
INCORRECT	CORRECT (STOP)	34 [30, 38]	90 [88, 92]
↓			
Question #3 ; What year are we in ?			
↓	↓		
INCORRECT	CORRECT (STOP)	13 [10, 16]	100 [99, 100]
↓			
Question #4 ; No further statistically significant splits (i.e. questions).			
↓			
Go on for more detailed cognitive evaluation			

APPENDIX L (cont.) ; Recursive Partitioning Strategy # 4B

4. Always pick the question with the highest NPV (NPV max)

(B) Only those who pass (CORRECT answer) the previous question go on to the next question. Those who answer a question incorrectly are considered cognitively impaired and are automatically required to go for further testing (indicated by 'STOP - further testing').

		Sensitivity (%) [95% CI]	Specificity
Question #1 ; Spell WORLD backwards.			
↓	↓		
INCORRECT (STOP - further testing)	CORRECT	61 [56, 65]	70 [66, 73]
	↓		
Question #2 ; What is today's date ?			
↓	↓		
INCORRECT (STOP - further testing)	CORRECT	79 [76, 83]	57 [53, 61]
	↓		
Question #3 ; Identify this body part (point to knuckle).			
↓	↓		
INCORRECT (STOP - further testing)	CORRECT	**84 [81, 87]**	**50 [46, 54]**
	↓		
Question #4 ; Three word recall (requires prior registration).			
↓	↓		
INCORRECT (STOP - further testing)	CORRECT	98 [97, 99]	13 [11, 16]

APPENDIX L (cont.) ; Recursive Partitioning Strategy #5A

5. Always pick the most Accurate question (Accuracy max)

(A) Only those who fail (INCORRECT answer) the previous question go on to the next question. Those who answer a question correctly are considered cognitively normal and do not require further testing (indicated by 'STOP').

Question #1 ; What is today's date (date of the month) ?		**Sensitivity (%) [95% CI]**	**Specificity**
↓ INCORRECT	↓ CORRECT (STOP)	53 [48, 57]	78 [74, 81]

↓

Question #2 ; Count backwards from 5.

↓ INCORRECT	↓ CORRECT (STOP)	34 [30, 38]	90 [88, 92]

↓

Question #3 ; What year are we in ?

↓ INCORRECT	↓ CORRECT (STOP)	13 [10, 16]	100 [99, 100]

↓

Question #4 ; No further statistically significant splits (i.e. questions).

↓

Go on for more detailed cognitive evaluation

APPENDIX L (cont.) ; Recursive Partitioning Strategy # 5B

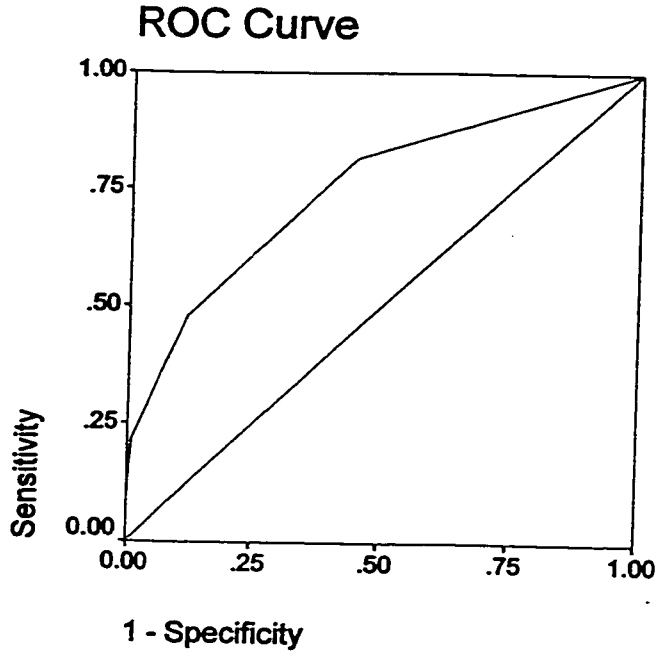
5. Always pick the most Accurate question (Accuracy max)

(B) Only those who pass (CORRECT answer) the previous question go on to the next question. Those who answer a question incorrectly are considered cognitively impaired and are automatically required to go for further testing (indicated by 'STOP - further testing').

Question #1 ; What is today's date (date of the month) ?		**Sensitivity (%) [95% CI]**	**Specificity**
INCORRECT (STOP - further testing)	CORRECT	53 [48, 57]	78 [74, 81]
Question #2 ; Spell WORLD backwards.			
INCORRECT (STOP - further testing)	CORRECT	79 [76, 83]	57 [53, 61]
Question #3 ; What year are we in ?			
INCORRECT (STOP - further testing)	CORRECT	**81 [77, 84]**	**56 [52, 60]**
Question #4 ; What country are we in ?			
INCORRECT (STOP - further testing)	CORRECT	**82 [79, 85]**	**54 [50, 58]**

APPENDIX M

ROC CURVE ANALYSIS of SET 4-1



Diagonal segments are produced by ties.

Area Under the Curve

Test Result Variable(s): SET4.1

Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
.754	.014	.000	.725	.782

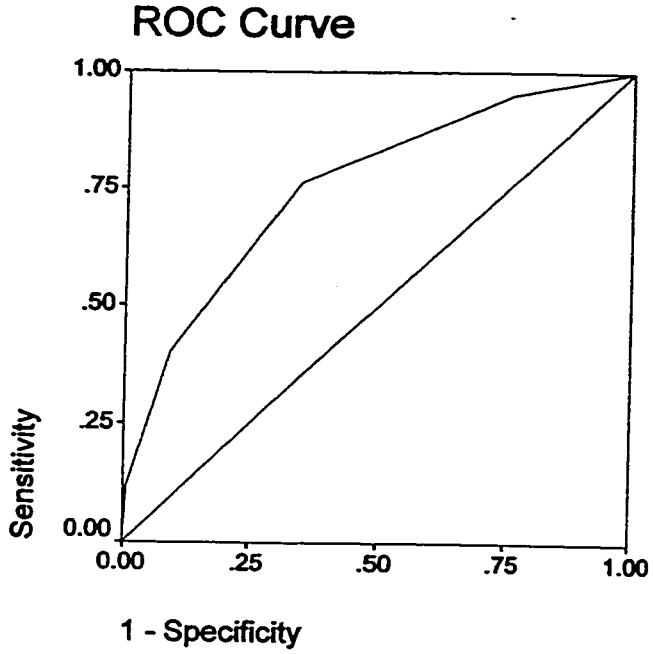
The test result variable(s): SET4.1 has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

APPENDIX M (cont.)

ROC CURVE ANALYSIS of SET 4-2



Diagonal segments are produced by ties.

Area Under the Curve

Test Result Variable(s): SET4.2

Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
.759	.014	.000	.731	.786

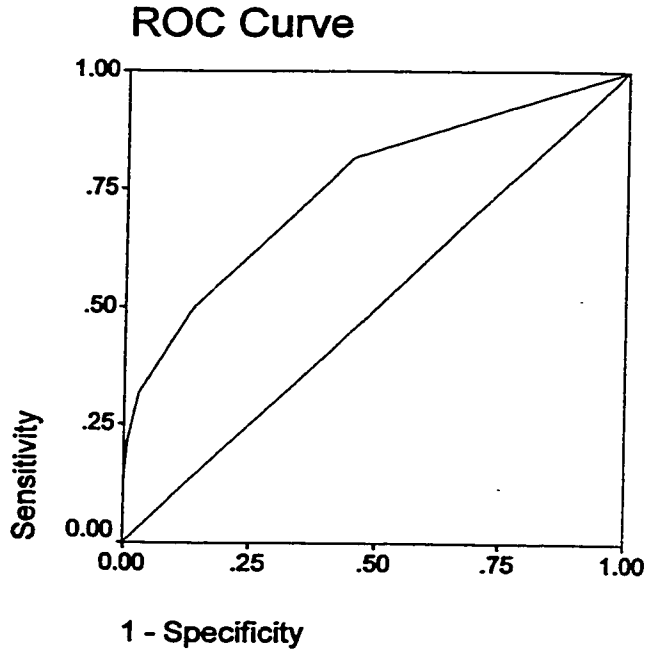
The test result variable(s): SET4.2 has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

APPENDIX M (cont.)

ROC CURVE ANALYSIS of SET 4-3



Diagonal segments are produced by ties.

Area Under the Curve

Test Result Variable(s): SET4.3

Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
.756	.014	.000	.728	.784

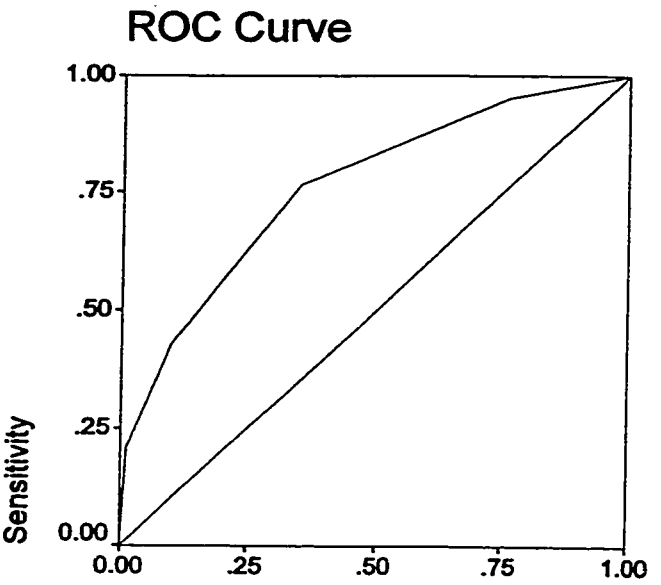
The test result variable(s): SET4.3 has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

APPENDIX M (cont.)

ROC CURVE ANALYSIS of SET 4-4



Diagonal segments are produced by ties.

Area Under the Curve

Test Result Variable(s): SET4.4

Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
.764	.014	.000	.737	.792

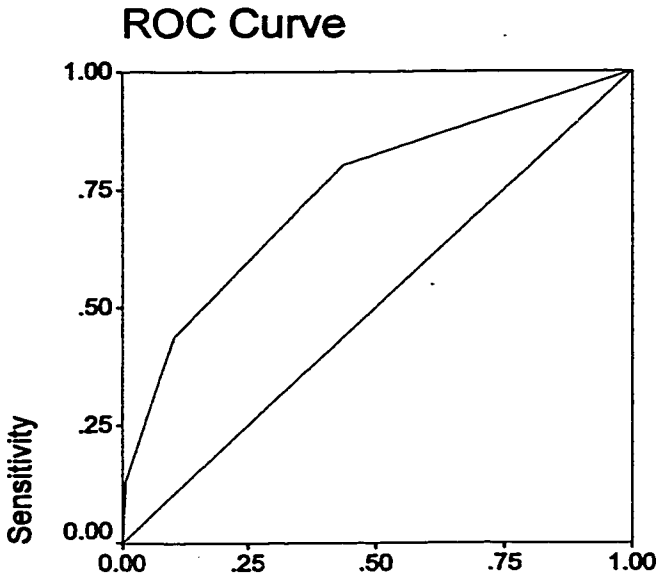
The test result variable(s): SET4.4 has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

APPENDIX M (cont.)

ROC CURVE ANALYSIS of SET 3-1



Diagonal segments are produced by ties.

Area Under the Curve

Test Result Variable(s): SET3.1

Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
.742	.015	.000	.713	.771

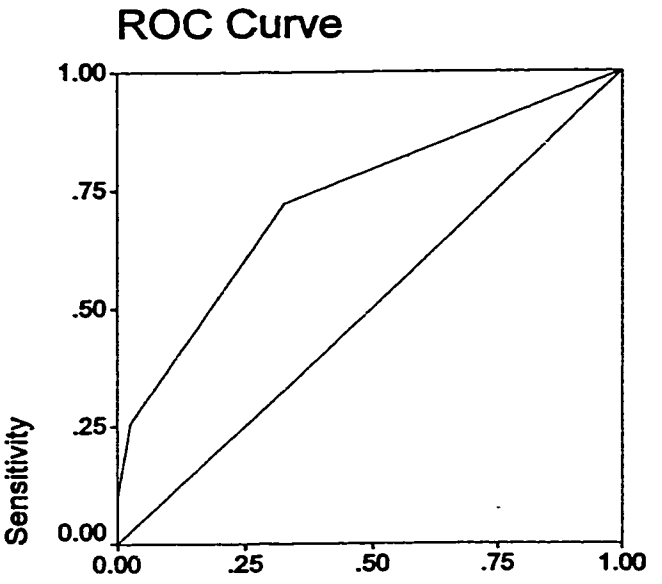
The test result variable(s): SET3.1 has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

APPENDIX M (cont.)

ROC CURVE ANALYSIS of SET 3-2



Diagonal segments are produced by ties.

Area Under the Curve

Test Result Variable(s): SET3.2

Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
.731	.015	.000	.702	.761

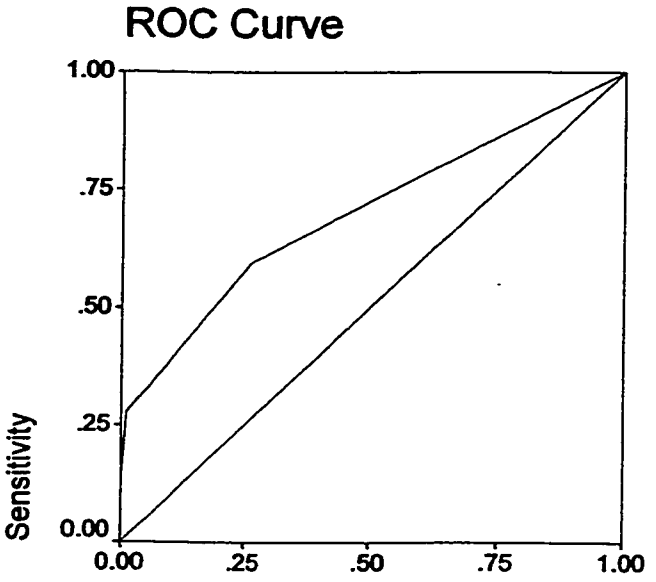
The test result variable(s): SET3.2 has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

APPENDIX M (cont.)

ROC CURVE ANALYSIS of SET 3-3



Diagonal segments are produced by ties.

Area Under the Curve

Test Result Variable(s): SET3.3

Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
.700	.016	.000	.669	.730

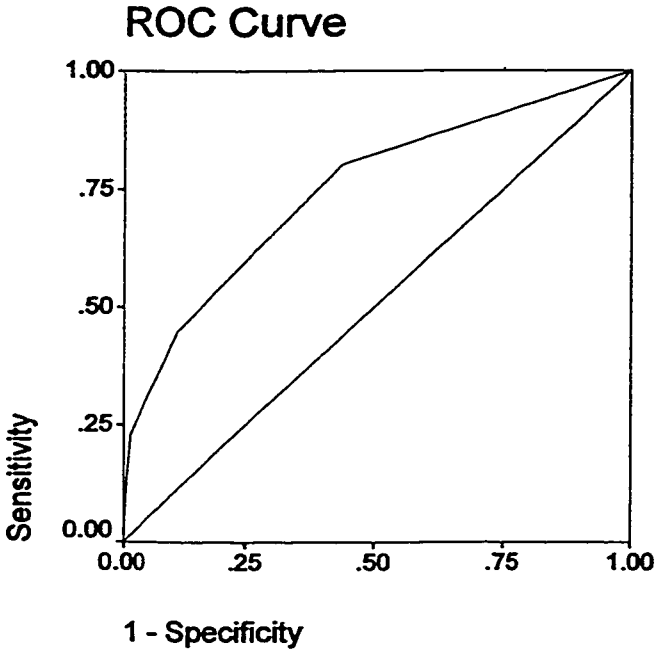
The test result variable(s): SET3.3 has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

APPENDIX M (cont.)

ROC CURVE ANALYSIS of SET 3-4



Diagonal segments are produced by ties.

Area Under the Curve

Test Result Variable(s): SET3.4

Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
.746	.015	.000	.717	.774

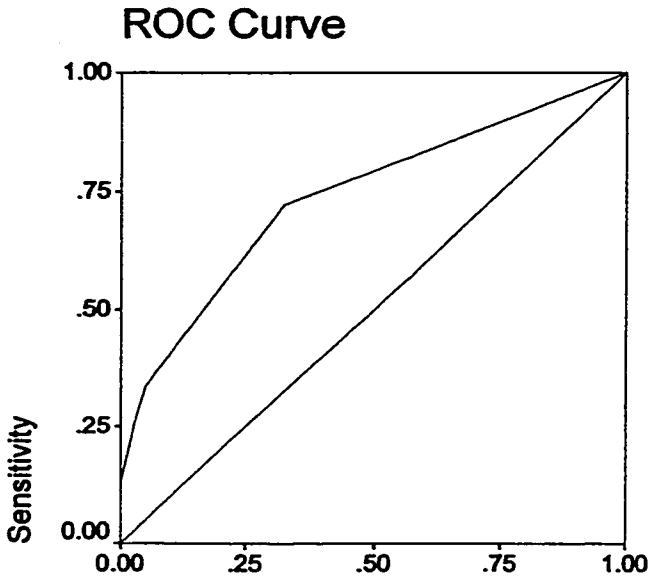
The test result variable(s): SET3.4 has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

APPENDIX M (cont.)

ROC CURVE ANALYSIS of SET 3-5



Diagonal segments are produced by ties.

Area Under the Curve

Test Result Variable(s): SET3.5

Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
.738	.015	.000	.709	.768

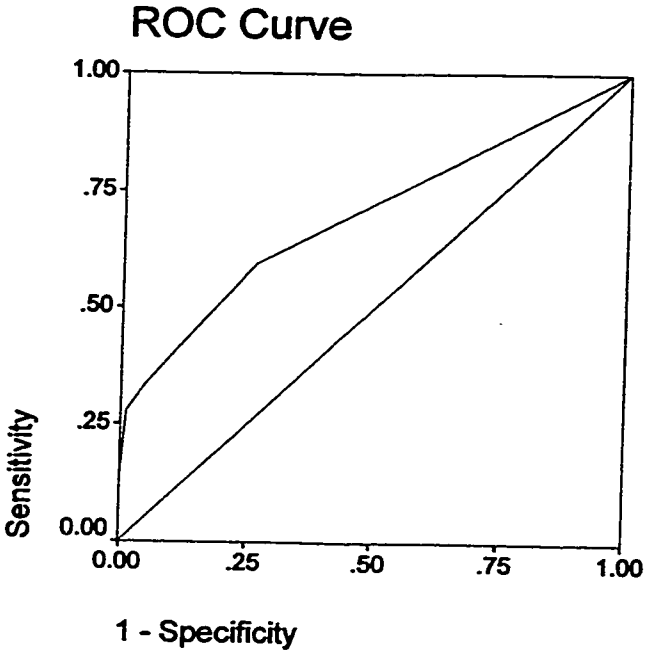
The test result variable(s): SET3.5 has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

APPENDIX M (cont.)

ROC CURVE ANALYSIS of SET 3-6



Diagonal segments are produced by ties.

Area Under the Curve

Test Result Variable(s): SET3.6

Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
.701	.016	.000	.670	.732

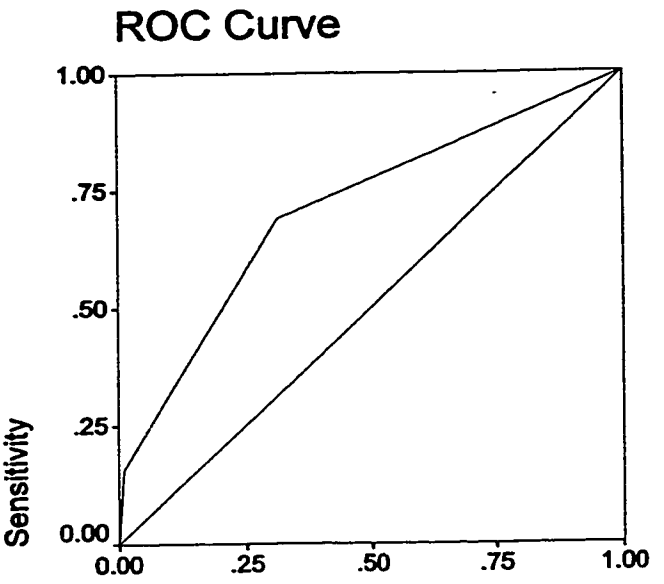
The test result variable(s): SET3.6 has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

APPENDIX M (cont.)

ROC CURVE ANALYSIS of SET 2-1



Diagonal segments are produced by ties.

Area Under the Curve

Test Result Variable(s): SET2.1

Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
.709	.015	.000	.679	.739

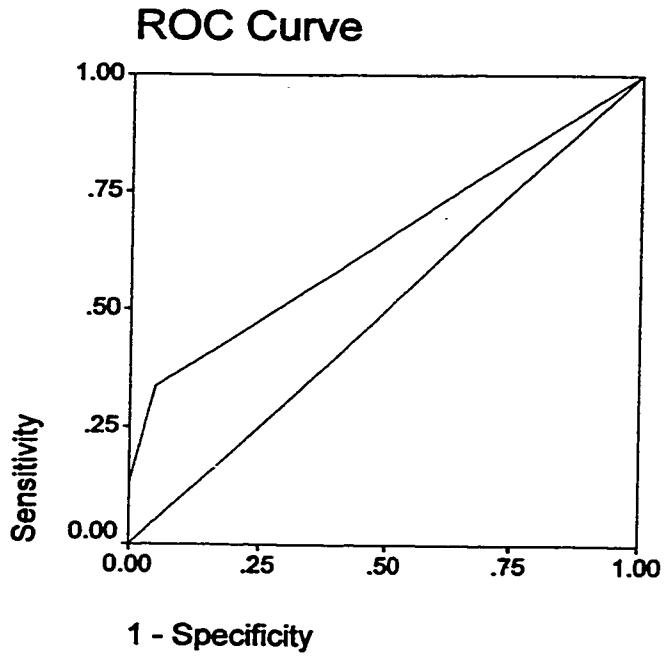
The test result variable(s): SET2.1 has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

APPENDIX M (cont.)

ROC CURVE ANALYSIS of SET 2-2



Diagonal segments are produced by ties.

Area Under the Curve

Test Result Variable(s): SET2.2

Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
.647	.017	.000	.614	.679

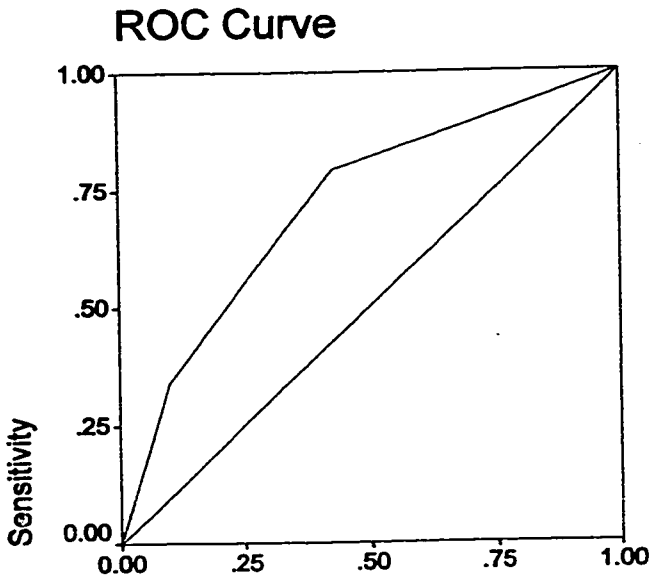
The test result variable(s): SET2.2 has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

APPENDIX M (cont.)

ROC CURVE ANALYSIS of SET 2-3



Diagonal segments are produced by ties.

Area Under the Curve

Test Result Variable(s): SET2.3

Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
.716	.015	.000	.686	.746

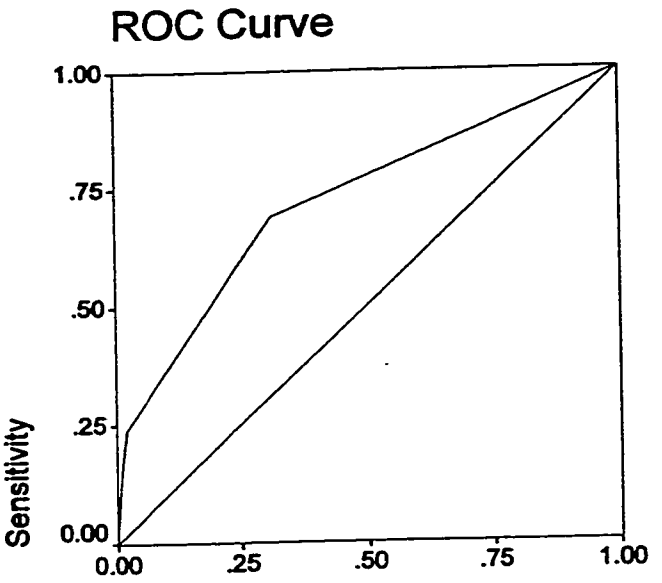
The test result variable(s): SET2.3 has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

APPENDIX M (cont.)

ROC CURVE ANALYSIS of SET 2-4



Diagonal segments are produced by ties.

Area Under the Curve

Test Result Variable(s): SET2.4

Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
.719	.015	.000	.689	.749

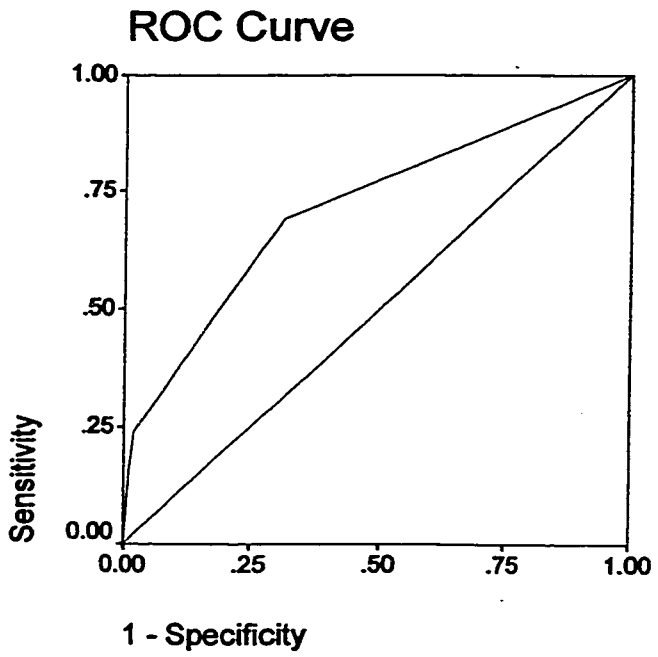
The test result variable(s): SET2.4 has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

APPENDIX M (cont.)

ROC CURVE ANALYSIS of SET 2-5



Diagonal segments are produced by ties.

Area Under the Curve

Test Result Variable(s): SET2.5

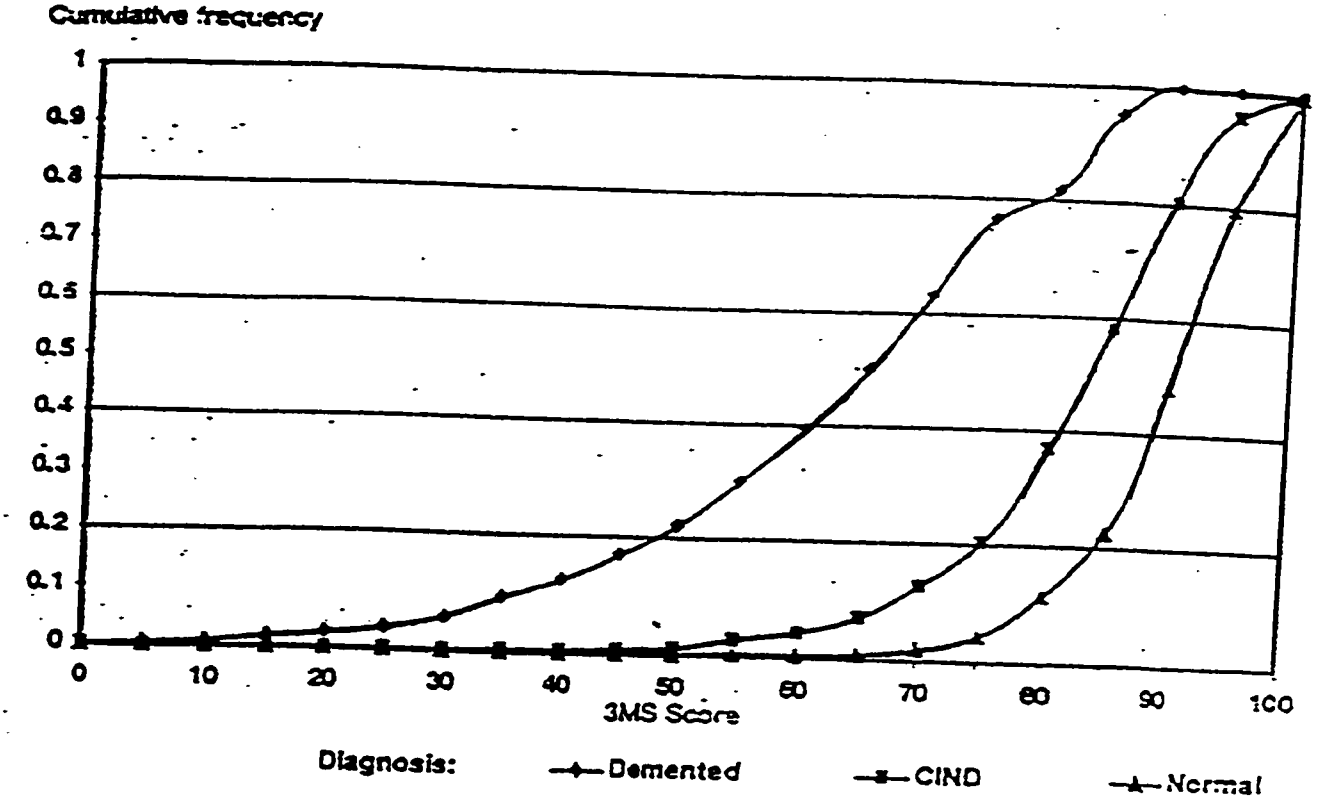
Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
.719	.015	.000	.689	.749

The test result variable(s): SET2.5 has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

APPENDIX N - Cumulative frequency distributions of 3MS screening scores for dementia, CIND (cognitive impairment no dementia), and cognitive normality (CSHA - 1)



APPENDIX O - Future Directions / Analyses

Future studies might include the following:

1. Further exploration of the process of selection of threshold or cut-off scores in the field of cognitive assessment. This appears to be a fertile area which has received insufficient study to date. Most of the analysis has been mathematical with little input from primary care clinicians. The decision regarding the relative weights of false positives vs. false negatives is essentially a clinical decision which will require the input from experienced primary care clinicians as well as experts in the field of cognitive impairment. This is likely a "moving target" as the relative weights may change as new therapies for dementia emerge.

2. Combination testing: repeating analyses such as those employed in this study but incorporating variables from the history, the physical examination, as well as the cognitive tests. This is essentially what is done when a full clinical cognitive assessment is performed. Later results of genetic studies, CT/MRI/PET scans might be incorporated into a multi-parameter clinical decision rule.

3. Continue to explore the utility of three level categorization (i.e. high probability, indeterminate [or intermediate] probability, and low probability of cognitive impairment). At present it is unclear whether this is a clinically useful approach as clinicians would tend to collapse the high and intermediate probability groups and subject them to the same in-depth cognitive testing and metabolic work-up. Such an approach might prove useful in the selection of patients who require more expensive or difficult to obtain procedures such as SPECT or PET scans (in the future).

4. If a screen developed in this study does become incorporated into clinical practice then we would consider following the "steps in developing a Clinical Decision Rule" as outlined by Stiell and Wells (depicted in Appendix Q) [27].

This might involve :
1. Prospective validation
2. Reliability studies
   Note; these will be difficult to interpret due to the fluctuation in cognition which defines the delirious subset of cognitively impaired persons.
3. Assessment of clinicians' interpretation of the results of the screen
4. Refinement of the screening tool
5. Assessment of the dissemination of the rule
6. Cost-effectiveness studies.

5. Explore strategies for "ruling in" vs. "ruling out" disease.

APPENDIX O - Future Directions / Analyses (cont.)

6. Explore strategies for sequencing highly sensitive tests (with high false positive rates) with highly specific tests (low false positive rates) given only to those who fail the first test. This parallels what is done in HIV testing. This approach has been suggested by Baldessarini and Finklestein [48]. Recursive Partitioning appears to be suited to this type of analysis.

7. Apply all of the above analyses to more specific diagnostic groups such as early dementia, delirium, Alzheimer's dementia, Lewy body disease, vascular dementia, etc..

8. Further exploration of the discriminating ability of questionnaires based on the observations by informants of behavioural changes over a given time period. Such questions are often very sensitive to early cognitive loss because they focus on change in comparison to an earlier baseline (i.e. a longitudinal perspective). This may be superior to mental status testing due to the fact that such testing cannot factor in the patient's natural baseline function (i.e. it has a cross-sectional perspective). The questionnaires essentially measure at two points in time while the mental status testing is limited to a single point in time. The questionnaires parallel the history taking techniques employed by clinicians with extensive experience in the evaluation of cognitive function. Unfortunately this approach requires the presence of a reliable caregiver and cannot be applied to unaccompanied patients.

9. It may be possible to develop simple cognitive screens, such as those developed in this thesis, which can be applied routinely every 6 to 12 months and which will be very sensitive to early change (i.e. longitudinal cognitive monitoring). Such screens would require a high degree of test-retest reliability.

10. Perhaps more sensitive and specific cognitive questions can be derived from existing detailed neuropsychological batteries (rather than the limited number of questions available in the MMSE or the 3MS). Such detailed testing is only performed in memory disorder clinics. Due to the different spectrum of disease, such tests would definitely require prospective validation in the intended clinical setting.

11. This thesis focussed on mild and moderate cognitive impairment as patients at both the levels are often missed. Given the increasing need to detect cognitive loss as early as possible in order to initiate treatment and prolong independence, it will become increasingly important to develop screens that are highly sensitive to early cognitive loss. Approaches 8 - 10, described above, may lead to the development of tools to detect early cognitive loss.

12. Acceptability research designed to study which screening tool formats (i.e. length of screen, types of questions, scoring....) are most appropriate and acceptable in which settings. This would also involve studying whether different tools are acceptable to different professions (i.e. nursing, physicians....). Internet based primary care research is developing (e.g. CARE) and may prove to be an ideal venue for studies of acceptability of various screens in primary care practice. It will be important to measure not only what clinicians say they would employ (this is affected by social desirability bias) but also to track the actual use of a tool in clinical practice.

In previous research, patients and families have demonstrated reluctance to permit cognitive testing due to a fear of precipitating a depression or a catastrophic reaction. Research into the form of testing which is most acceptable to patients and families is also required.

13. Reliability Research - studies of inter-rater and intra-rater (test-retest) reliability are required for all new cognitive screens. One cannot assume reliability of shorter screens because their questions were drawn from larger tests with established reliability. It is possible that the least reliable questions might be selected for the shorter test. Consequently, it is necessary to repeat reliability studies for all new screens. Alternatively one might test reliability of each individual question in the larger test and then employ a minimal reliability threshold as a selection criteria prior to selection of questions for the shorter test (i.e. prior to multivariate analysis).

14. A more detailed assessment approach, based on recursive partitioning, could also be designed for specialty clinics evaluating cognitive loss. Such clinics would include memory disorder clinics and geriatric day hospitals. The approach presently employed in these clinics is similar to the branching approach of recursive partitioning but has not been derived empirically.

15. A study of the psychometric properties of the Clock Drawing as it is actually employed in clinical practice (i.e. the "eyeball method" of scoring rather that the formal scoring techniques promoted by published studies). This study is presently in the planning stage under the direction of the SCOHS Primary Care of the Elderly research group.

16. Further exploration of the utility of recursive partitioning in this field. The decision to use the same criteria for each question in an algorithm (e.g. always ask the most specific question) may be too restrictive. By varying the criteria at each step one may develop a series of questions with more useful psychometric properties. This approach is supported by Dr. Stiell. Unfortunately this would result in a loss of reproducibility of the technique (i.e. objective #2 of this thesis) and consequently did not form part of this study.

17. Further exploration of the classification (impaired vs. normal) and the probabilistic (probability of impairment) approaches. These approaches may be useful to clinicians with different levels of expertise and clinical focus.

APPENDIX O - Future Directions / Analyses (cont.)

18. A great deal of qualitative information (e.g. speed and ease of response) is intuitively employed by physicians and psychologists with experience in the evaluation of cognition. This qualitative information may be important in the evaluation of early cognitive loss. Unfortunately such data is lost when simple cognitive screens are used. It may prove useful to attempt to convert such qualitative observations into reliable standardized quantitative data.

19. Perhaps the nihilism regarding the clinical utility of multivariate equations is not completely justified. As mobile computerized devices (e.g. Palm Pilots or Portable PCs) become a part of physicians' everyday environment, we may reach a point where multivariate equations may be programmed into such devices and employed routinely in clinical practice. While programmable calculators have been available for years, few MDs carry such calculators around and fewer still know how to program complex equations into such devices.

20. Derive or validate/refine similar clinical decision rules for other settings such as acute care and nursing homes. All of the above future strategies can also be applied to these settings.

21. The studies listed above should be done in both Anglophone and Francophone environments. Further development of cognitive screens for other cultural groups should be considered. All of the above future strategies can also be applied to these groups.

22. Validation of the equations employed to correct for verification bias to determine which equation is most predictive of actual test performance in unselected patients.

APPENDIX P - Expanded set of screens which merit prospective validation
(Additional screens are highlighted)

[1] Single Question	χ^2	P-value	Sens(%)[95% CI]	Spec(%)[95% CI]
Three word recall	47.19	0.0000	**95 [93, 97]**	**18 [15, 22]**
Similarity (arm/leg)	33.15	0.0000	**93 [92, 96]**	**17 [14, 20]**

	Cut-point	Sensitivity (%) [95% CI]	Specificity (%)[95% CI]
Sets of 4 questions			
[2] Set 4-1	0/1	9 [7, 11]	100 [99, 100]
	1/2	22 [18, 25]	99 [98, 100]
	2/3	48 [44, 52]	88 [85, 90]
	3/4	**82 [78, 85]**	**55 [51, 59]**
[3] Set 4-2	0/1	12 [9, 15]	99 [98, 100]
	1/2	40 [36, 44]	91 [88, 93]
	2/3	76 [72, 80]	65 [61, 69]
	3/4	**95 [93, 97]**	**24 [21, 28]**
Sets of 3 questions			
[4] Set 3-1	0/1	13 [10, 16]	99 [98, 100]
	1/2	44 [40, 48]	89 [87, 92]
	2/3	**80 [77, 83]**	**56 [52, 60]**

APPENDIX P cont

Expanded set of screens which merit prospective validation

[5] ***Recursive Partitioning strategy 1B and 4B*** **(add one more question)**
Always pick the most Sensitive question (Sens max)

(B) Only those who pass (CORRECT answer) the previous question go on to the next question. Those who answer a question incorrectly are considered cognitively impaired and are automatically required to go for further testing (indicated by 'STOP- further testing').

		Sensitivity (%) [95% CI]	Specificity
Question #1 ; Spell WORLD backwards.			
INCORRECT (STOP - further testing)	CORRECT	61 [56, 65]	70 [66, 73]
Question #2 ; What is today's date ?			
INCORRECT (STOP - Further testing)	CORRECT	79 [76, 83]	57 [53, 61]
Question #3 ; What is this called (point to knuckle) ?			
INCORRECT (STOP - Further testing)	CORRECT	**84 [81, 87]**	**50 [46, 54]**
Question #4 ; Three word recall (requires prior registration)			
INCORRECT (STOP - further testing)	CORRECT	**98 [97, 99]**	**13 [11, 16]**

APPENDIX P cont.

Expanded set of screens which merit prospective validation

[6] Recursive Partitioning Strategy # 5B

Always pick the most Accurate question (Accuracy max)

(B) Only those who pass (CORRECT answer) the previous question go on to the next question. Those who answer a question incorrectly are considered cognitively impaired and are automatically required to go for further testing (indicated by 'STOP - further testing').

Question #1 ; What is today's date (date of the month) ?		**Sensitivity (%) [95% CI]**	**Specificity**
INCORRECT (STOP - further testing)	CORRECT	53 [48, 57]	78 [74, 81]
Question #2 ; Spell WORLD backwards.			
INCORRECT (STOP - further testing)	CORRECT	79 [76, 83]	57 [53, 61]
Question #3 ; What year are we in ?			
INCORRECT (STOP - further testing)	CORRECT	**81 [77, 84]**	**56 [52, 60]**

APPENDIX Q: Checklist of standards for 6 stages in the development of a clinical decision rule. (The italicized and underlined sections have been addressed in this study.)

1. Is there a need for the decision rule ?

Prevalence of the clinical condition.
Current use of the diagnostic test.
Variation in practice.
Attitudes of physicians.
Clinical accuracy of physicians.

2. Was the rule derived according to methodologic standards ?

Definition of outcome.
Definition of predictor variables.
Reliability of predictor variables.
Selection of subjects.
Sample size.
Mathematical techniques.
Sensibility of the decision rule
Accuracy.

3. Has the rule been prospectively validated and refined ?

Prospective validation.
Selection of subjects.
Application of the rule.
Outcomes
Accuracy of the rule.
Reliability of the rule.
Physicians' interpretation.
Refinement.
Potential effect.

4. Has the rule been successfully implemented into clinical practice ?

Clinical Trial.
Effect on use.
Acceptability.

5. Would use of the rule be cost-effective ?

6. How will the rule be disseminated and implemented ?