ることがことができる．このため研究者は，過去の研究結果などを参照し，十分な検出力を期待できる標本数をあらかじめ設定することが望まれる．たとえば脳梗塞患者と健常対照者の血清コレステロール値の差を検定する場合，血清コレステロール値の標準偏差が20mg/dlで両群間の差が5mg/dlであれば，80％の検出力で5％の危険率による有意な差が検出される，すなわち80％の確率で，差がないとする帰無仮説が危険率5％で棄却されるためには，各群約250名ずつの標本数が必要となり，両群間の差が2mg/dlであれば，各群約1,500名の標本数が必要となるのである．

### ●2．有意水準が小さいほど差が大きい

この誤りはたとえば高コレステロール血症の治療薬についてある論文でAという薬が3％の危険率で，また違う論文ではBという薬が0.01％の危険率で血清コレステロールの値を減らすと報告していた場合に，Bという薬のほうが，Aという薬よりよく効くと解釈する誤りである．このことも上述の精度の問題であり，標本数が大きな影響を及ぼす．たとえば平均血清コレステロール値が310mg/dl，標準偏差が36mg/dlの群で，ある薬物が1ヵ月で平均4mg/dl，標準偏差8mg/dlだけ血清コレステロール値を下げた場合，対象数が25名である場合と，その4倍である場合ではこれらの数値はまったく変わらないにもかかわらず危険率は2.2％から0.0003％と大きく変わってくる．しかし，薬物の効果はまったく差がないのである．この例における薬物の効果のような研究者が研究で捉えたいと思う効果（差）の強さを効果量（effect size）という．危険率は第1種の過誤を生じる確率のことであり，第1種の過誤は第2種の過誤と同様，偶然誤差により生じる．したがって危険率の値は効果量に加え，標本数の影響を大きく受けるため，効果量の代わりとして解釈してはならないのである．

### ●3．多重比較・反復比較による「有意な差」の検出

多重比較とは，たとえば脳梗塞により生じた麻痺に対する治療薬の効果を調べるときに，プラセボ投与群，5mg投与群，10mg投与群，20mg投与群の4群で，効果を，プラセボ投与群vs 5mg投与群，5mg投与群vs 10mg投与群など，2群ずつ総当たりでくり返し検討することであり，反復比較とはたとえば脳梗塞により生じた麻痺に対する治療薬の効果を調べるときに顔面筋力，上肢筋力，下肢筋力，上肢筋持久力，下肢筋持久力を左右で検定するなど同じような項目をくり返し検定することである．このことはすでに1985年にGodfrey[1]により『N Engl J Med』誌で警告されているにもかかわらず，現在でも論文の査読などをしているとよくみる誤りである．前回述べたように，これらの場合に，たとえば多重比較ではたとえばプラセボ投与群と20mg投与群のあいだで，反復比較ではたとえば左の上肢筋力で「有意な」相違が認められたからといって，その項目で効果があると短絡的に解釈してはならない．上述のように危険率とは実際に帰無仮説，この例の場合はこの治療薬が脳梗塞により生じた麻痺の改善に効果がない，が正しくてもその確率で偶然に差が認められる確率のことである．したがって5％の危険率があるということは20回に1回はそのような偶然による差が生じうることを意味している．このため単純に考えると20回同じような検定をくり返せば1つは「有意な」相違が認められることになる．このように同じような統計検定のくり返しは第1種の過誤の危険を増大させる．このことを防ぐために研究者は，前回述べたように同じような検定のくり返しを避けるため主となるエンドポイントを設定するなどの研究計画の工夫や，分散分析と適切なpost hoc検定，あるいはBonferroni補正のような適切な統計学的方法の選択をすることが望まれる．

### ●4．統計学的関連があれば因果関係がある

この誤りは厳密な意味では統計学的な解釈の誤りではないが，まさに，統計結果の解釈の誤りである．たとえば脳梗塞患者と健常対照者で収縮期血圧を比較したところ，脳梗塞患者で収縮期血圧が有意に高かった場合，収縮期血圧の高値は脳梗塞の原因であると解釈する誤りである．真に因果関係がある場合，すなわち真に収縮期血圧の高値が脳梗塞の原因の一つである場合にも，このような統計学的関連がみられるが，その他にも，脳梗塞により収縮期血圧が上昇するなど因果関係が逆転している場合，収縮期血圧と関連のある他の要因，たとえば年齢が真の原因である場合，さらには第1種の過誤により偶然に統計学的関連がみられる場合などが考えられる．因果関係を判断するためには，単に統計学的に関連があることだけではなく，時間的

に順序関係がある，関連が強固で普遍的である，関連が用量反応的であるなど種々の条件を考慮しなくてはならない．

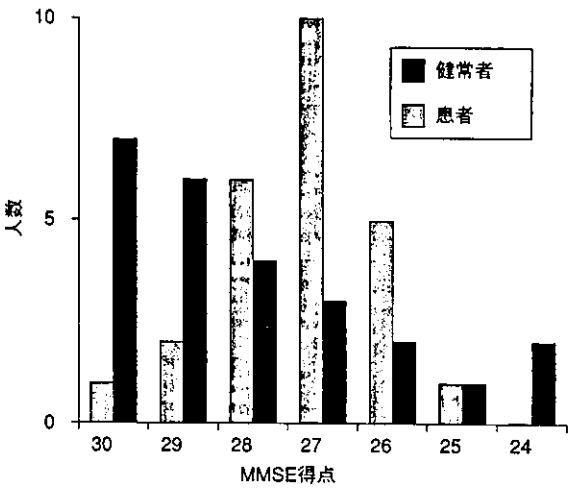
## II 誤った統計解析の方法の選択

統計解析の方法はアウトカムの変数の尺度（間隔尺度，順序尺度，名義尺度），変数の分布の型，変数の数などにより適切に選択されなければならない．むしろ研究者は，適切な統計解析の方法が適用できるようにアウトカムを設定しデータを収集するように研究計画をつくらなければならないといって過言ではないであろう．誤った統計学的方法を適応すると誤った結果を導き出す原因となる．ここでは臨床研究でよくみられる誤りについて述べる．

### ●1．パラメトリック手法とノンパラメトリック手法

パラメトリック手法およびノンパラメトリック手法の厳密な定義はむずかしいが，一般的にはパラメトリック手法は母集団に特定の分布の型を要求する手法，ノンパラメトリック手法は母集団の分布によらない（distribution free）手法であると考えて大きな間違いはない．ここでは詳細に述べることはできないが，パラメトリック手法にはStudentのｔ検定や分散分析などが，ノンパラメトリック手法にはMann-Whitney U検定やKruskal-Wallis検定などがある．統計解析方法の誤用としてよくみられるものの第一は，パラメトリック手法を用いてはならないデータに対しパラメトリック手法を用いるものである．

パラメトリック手法を適応できるのは，アウトカムのデータが1. 計量値であること，2. 間隔尺度であること，3. 母集団が正規分布であることが仮定されること，4. 比較される各群の分散が等しいことの4条件を満たす場合のみである．したがって，喫煙者と非喫煙者の脳梗塞の罹患頻度を比較する場合（計数値）や，脳梗塞患者と健常者で，喫煙数を1. 全くすわない，2. 1日1～10本，3. 1日11～20本，4. 1日21本以上というような順序カテゴリー尺度で評価し比較する場合はStudentのt検定を用いてはならない．さらに，計量値であり間隔尺度である場合でも，母集団の正規分布が仮定できない場合にはパラメトリック手法を用いてはならない．母集団の分布がわからない場合には，グラフの使用が直感的で有用である．たとえばラクナ梗塞の患者と健常者の認知機能の差を認知機能スクリーニング検査であるMini-Mental State Examination（MMSE）の成績を用いて検討したときに図❶のような結果になったとする．健常者の成績は平均28.0点，標準偏差1.9，患者の成績は平均27.2点，標準偏差1.1点である．この2群をStudentのｔ検定を用い比較すると危険率6.2％で差がないとする帰無仮説を棄却することができない．しかし，よくグラフをみると明らかなように健常者のMMSEの成績は正規分布をしているとはとてもいえない．これは健常者のMMSEの成績分布には30点満点を超える点数をとれないことにより生じる天井効果があるためである．このため母集団自体に正規分布を仮定することが誤りであると考えられる．そこでノンパラメトリック手法であるMann-Whitney U検定を用い比較すると危険率2.7％で帰無仮説を棄却できる．一般にパラメトリック手法は，ノンパラメトリック手法にくらべて検出力が高いといわれているが，このように誤用をおこなった場合にはそのかぎりではない．各群の分散が等しいといえない場合も原則としてパラメトリック手法を用いるべきではないが，ｔ検定におけるWelchの変法のように分散が異なる場合にも適切に用いることのできる方法が用意されている場合もある．




**図❶ ラクナ梗塞の患者と健常者の認知機能の差の一例**

### ●2．対応のある場合とない場合

対応のある検定とは，たとえば治療前と治療後の比較

のように同一の個体についてくり返しデータが測定されている場合の検定をいう．統計解析方法の誤用としてよくみられるものの第二は対応がある場合に対応がない場合の検定を用いるものである．対応のある検定手法には，対応のあるｔ検定，くり返しのある分散分析，対応のあるWilcoxon検定などがあるが，よく誤りがみられるものに対応のある計数データに対しMcNemar検定ではなくカイ自乗検定を用いているものがあげられる．たとえば同じ100名のラクナ梗塞患者で，X線CTでは73名に，MRIでは83名にラクナ梗塞が検出できた場合に，この2つの検査の診断力に差があるかを検定するのに73-27-83-17の四分表を作成し（表❶a），カイ自乗検定を適応するのがこの誤りである．この四分表を作成すると，症例数は合計200名になり，実際の症例数100名とは異なる．この場合は両方でラクナ梗塞が検出できた人数（たとえば69名とする），X線CTのみで検出できた人数（4名），MRIのみで検出できた人数（14名），両方ともで検出できなかった人数（13名）の四分表を作成し（表❶b），X線CTのみで検出できた人数とMRIのみで検出できた人数とを比較検討しなくてはならない．この手法がMcNemar検定である．表❶aのカイ自乗検定をおこなうと危険率は8.5％であるが，表❶bのMcNemar検定をおこなうと危険率は3.4％となり，結果が異なってくる．

## ●3．交絡因子

検討しようとする統計学的効果に影響を与える他の因子を交絡因子confounding factorとよぶ．たとえば脳梗塞の発症率に対する喫煙の影響を検討する場合，喫煙の他にも年齢や高血圧，糖尿病などの因子が脳梗塞の発症と関連していると考えられる．これらの影響を考慮しなければ，喫煙の脳梗塞の発症率に対する影響を検討することはできない．たとえばある研究で喫煙者は非喫煙者にくらべ有意に脳梗塞の発症率が高かったという結果が得られたとする．しかし，もし喫煙者のほうが年齢が高かったり，高血圧患者が多かったりした場合は，この喫煙の有無によると思われた差は，実は年齢の差や高血圧の有無による差によるものかも知れないからである．第三の誤用はこのような場合にこれらの交絡因子の影響を検討する適切な手法をとらずに，喫煙による効果であると速断してしまうものである．

交絡因子の影響を適切に除くことを交絡因子の調整あるいは補正（control）とよぶ．交絡因子の調整は，標本の選択段階でおこなっておくことが望ましい．完全な無作為確率的サンプリング法の採用，既知の交絡因子を有するものの除外（この例の場合は高血圧患者），さらには既知の交絡因子を一致させた症例対照対の作成などがこれにあたる．しかし，臨床研究ではこのような標本選択をおこなうことはむずかしく，さらにはそのような標本選択自体が極めて限定的な標本集団をつくり出し，

表❶　X線CTとMRIによるラクナ梗塞の検出率の比較

a)

| | X線CT | MRI | 計 |
|---|---|---|---|
| 検出(+) | 73 | 83 | 100 |
| 検出(-) | 27 | 17 | 100 |
| 計 | 100 | 100 | 200(?) |

b)

| | | MRI | | |
|---|---|---|---|---|
| | | 検出(+) | 検出(-) | 計 |
| X線CT | 検出(+) | 69 | 4 | 73 |
| | 検出(-) | 14 | 13 | 27 |
| | 計 | 83 | 17 | 100 |

表❷ 統計でごまかす10の方法（Greenhalgh T，1997[2]より引用，著者訳）

| |
|---|
| 1．全部のデータをコンピュータに入れて，$P<0.05$の有意なものを全部を報告しよう． |
| 2．もしベースラインでの差が介入群のほうに有利だったら，その差は補正しないでおこう． |
| 3．正規分布しているかどうか確かめないでおこう．もしそれをしたら，おもしろくないノンパラメトリックテストで行き詰まってしまうかも知れない． |
| 4．脱落例と無反応例はすべて無視しよう．そうすれば分析は完全に治療を受けた患者だけでできる． |
| 5．データとデータは必ず対応させてプロットできて，ピアソン相関係数rが計算できると考えよう．そして有意なrは因果関係を証明していると考えよう． |
| 6．もしはずれ値（グラフのなかで飛び離れているもの）が計算の邪魔になるのなら，それを消そう．でももしそのはずれ値が何かの助けになっているのなら，たとえそれが何かの間違いのようであったとしてもそのままにしておこう． |
| 7．もし信頼区間が2群間で差がないところを含んでいたらそれを報告しないでおこう．本文で簡単に述べてもいいけど，グラフに描かないで，結論では無視しよう． |
| 8．6ヵ月間の臨床試験で，2群の差が4.7ヵ月半で有意になったら臨床試験を止めて論文を書きはじめよう，逆にもし6ヵ月の時点でもう少しで有意だったら試験をあと3週間延ばそう． |
| 9．もし結果がおもしろくなかったら，コンピュータで他に何か違って動くサブグループがないかみてみよう．ついには52～61歳の中国人女性に効くということが見つかるかも知れない． |
| 10．もし予定していたデータ分析方法で望んでいた結果が得られなかったら，他の方法で計算してみよう． |

得られる結果の一般化を困難にしてしまうこともある．したがってデータ解析時に統計学的手法を工夫して，調整することが必要となる場合が多い．その方法としては交絡因子の有無により層化して検討する方法や，ロジスティック回帰分析などの多変量解析を用いる方法がある．

## おわりに

1997年，Greenhalgh[2]は，論文を書く際に「統計でごまかす10の方法」という英国人流のジョークを『BMJ』誌に発表している（表❷）．このなかの多くは標本の抽出，統計学的方法の選択，および統計学的解釈の諸問題として前回および今回述べたことである．これらを実践してみようという人，地でおこなったことのある人，ジョークと思えない人は要注意である．研究者は誤った研究結果を導き出さないために，これらの誤りが生じていないかどうか，つねに検討することが必要である．また臨床家が診療の根拠となる文献を検討する際にも，その研究が正しい統計学的方法を適用し，その結果を正しく解釈しているのかどうかを批判的に吟味する必要がある．同時に，研究者はそのような批判的な吟味が可能であるように適切な情報を開示した論文を発表することが望まれる．

●文　献●

1) Godfrey K：Statistics in practice. Comparing the means of several groups. *N Engl J Med* **313**：1450-1456, 1985
2) Greenhalgh T：How to read a paper. Statistics for the non-statistician. II："Significant" relations and their pitfalls. *BMJ* **315**：422-425, 1997

EBM と 医療統計

4

# アウトカムの尺度とエンドポイントの評価法

博野信次，森　悦朗*
HIRONO Nobutsugu & MORI Etsuro
愛媛大学医学部看護学科
*兵庫県立姫路循環器病センター

## 著者プロフィール

（ひろの・のぶつぐ）

1959年，大阪生まれ．
【略歴】1984年，大阪市立大学医学部卒業．同年，大阪市立大学附属病院第一内科学教室研修医，住友病院神経内科医員．1990年，馬場記念病院神経内科医員．1992年，兵庫県保健環境部県立病院局経営課脳研準備室主査．1993年，兵庫県立高齢者脳機能研究センター臨床研究科研究員．1998年，米国University of California Los Angeles,School of Medicine客員研究員．1999年，兵庫県立高齢者脳機能研究センター臨床研究科研究員．2002年，愛媛大学医学部看護学科教授．
【専門】臨床痴呆学，神経心理学．
【研究テーマ】痴呆の神経心理学，痴呆の臨床神経学．
【趣味・愛読書】古代中国史

## はじめに

この講座は4回の連載ということではじまり，今回が最終回の予定であったが，予想外に好評であるとのことで，あと1年間延長され8回の連載となった．そのため今回は，これまで紙数の都合で端折ってきた部分を急遽補完することにした．

臨床家が臨床研究を批判的に読み，また臨床研究家が臨床研究を実践する際には，その研究が，1．妥当な仮説にもとづいて，2．それを検証するよう適切に計画された研究デザインにより，3．適切に選択された対象者に対し，4．適切な評価法を用いてアウトカムが評価され，5．その結果を適切な統計方法を用いて解析しているか否かを検討しなくてはならない．そのなかで，研究の質に直接関係する研究デザインとその種類，バイアスを含む対象者選択の問題，および誤りやすい統計解析のポイントなどについては前回までに詳述してきたが，アウトカムの評価尺度の信頼性と妥当性，検査の診断力および各研究デザインにおける予測因子とエンドポイントとの関連の評価法などについては，やや説明が足りなかったと思われるので，ここで詳述する．話は臨床研究（臨床疫学）からEBMの技法に移ってきているが，これらのことは臨床研究のデザインをおこなううえでも当然ふまえておかなければならない基本的な事柄である．

## I　評価尺度の信頼性と妥当性

アウトカムの測定には，種々の評価尺度（スケール）が用いられる．臨床研究を吟味する際には，まず用いられた評価尺度の信頼性と妥当性について検討しておく必要がある．評価尺度の信頼性とは同一あるいは異なった計測者が測定をくり返したとき安定した結果が得られるかどうかということを意味し，妥当性とは測定すべきものを正当に計測しているかどうかということを意味する．たとえば，脳血管障害により患者に生じた日常生活活動（ADL）障害を評価するためにある評価尺度が使われていた場合，評価者が異なれば結果が異なったり，同じ評価者でもおこなうたびにその結果が異なったりしては評価があてにならなくなってしまう．臨床試験では，くり返し評価をおこない，複数の評価者がいるのが普通で，評価が安定しないことは致命的である．このため評価尺

度には，定められている方法通りにおこなえば，誰が評価しても，何回評価しても同じ結果が出ることが要求される．これが信頼性である．またADLを評価するために用いた評価尺度が，実際にはADLではなく，麻痺の程度や感覚障害の程度などの機能障害を評価していてはいけない．ADL評価尺度は正しくADLを評価していなくてはならないのである．これが妥当性である．信頼性と妥当性との相違は，たとえば緑日の射的を思っていただければわかりやすい．コルク鉄砲でお菓子などの的を撃ち，見事，お菓子が下に落ちればそのお菓子がもらえるという，あれである．鉄砲のなかには何度撃っても，誰が撃っても，たとえば10cm左外側に玉がいってしまうものがあったりする．この鉄砲の場合，必ず的の左外側10cmにいくわけであるから，当たる位置の一致性，すなわち信頼性は高い．しかし的には決して当たらないので妥当性は低いことになる．反対に，撃つたびにどちらにずれるかわからない鉄砲の場合は，たまにはまっすぐに飛び，的に当たるのであれば，必ずはずれる左ずれ鉄砲よりは妥当性が高いかもしれないが，当たる位置の一致性，信頼性はきわめて低いことになる．臨床家が研究論文を評価する場合には，そこで用いられている評価尺度が過去の研究により信頼性と妥当性が示されている確立されたものであるのか，あるいは新しいものである場合は，その研究のなかで信頼性と妥当性が適切に示されているのかどうかを吟味しておかなければならない．

## 1．信頼性

信頼性の評価は，異なった検者間での評価結果の一致性（検者間信頼性inter-rater reliability），あるいは同一検者あるいは異なった検者によりくり返しおこなわれた評価結果の一致性（検査—再検査信頼性test-retest reliability）を検討することによりおこなう．単に検者間信頼性といった場合には，ある検査をおこなっているときに，同時に2人以上の評価者が評価をおこなうものを意味する場合が多く，必ずしも異なった検者により，別々に検査自体をおこなうことを意味しないことに注意を要する．同様に単に検査—再検査信頼性といった場合には，同一の検者により評価されている場合が多い．

これらの異なった評価間での一致性を検討するのであるが，一致率が高ければ信頼性が高いと単純にいえるわけではない．たとえば，ある地域の住民100名の頭部

表❶ 2人の評価者による脳梗塞の診断の一致

| | | 評価者1 | | |
|---|---|---|---|---|
| | 脳梗塞 | あり | なし | 合計 |
| 評価者2 | あり | 3 | 7 | 10 |
| | なし | 7 | 83 | 90 |
| | 合計 | 10 | 90 | 100 |

MRIをもとに，ある評価方法により脳梗塞の有無を評価したとしよう．このとき，Aという評価者とBという評価者の評価結果が86％一致していれば，その評価法は信頼性が高いといえるだろうか．もし対象者の90％が健常であり，両方の評価者がそれぞれ90人を健常としていた場合には，偶然に両方の評価者が健常とする確率は90％×90％＝81％，脳梗塞とする確率は10％×10％＝1％となり，偶然の一致率は82％にもなる．このため，86％の一致率があったといっても偶然の一致率を大きく上回らず，この評価法が信頼性のあるものだとはいえない（表❶）．$\kappa$（カッパ）係数という一致率の尺度はこの偶然の一致を考慮したものである[1]．$\kappa$係数は実際の一致率から偶然の一致率を引いたものを，1から偶然の一致率を引いたもので割った値である．先ほどの例では（0.86－0.82）/（1－0.82）＝0.22となる．この$\kappa$係数が0.4～0.6で中等度の（moderate），0.6～0.8でかなりの（substantial），0.8～1.0でほぼ完全な（almost perfect）一致性があるとされている[2]．$\kappa$係数には3人以上の評価者間での一致性を見ることができるものや[3]，名義尺度ではなく順序尺度の場合，たとえば5段階評価の1と2の違いの場合と1と5の違いの場合を区別する重み付け$\kappa$係数がある[4]．また評価結果が間隔尺度の場合には階級内相関係数intra-class correlationを用いることが多い[5]．

信頼性評価のなかに通常含まれるものに，内的整合性（内的一貫性，内的一致性internal consistency）がある．これはある評価法のすべての項目が同じものを（たとえば記憶障害）を計測していることを示すものである．内的整合性の評価方法には「折半法」すなわち検査を任意に折半して相関係数を取る方法があり，考えうる折半方

法すべてを勘案したものがCronbachのα係数である[5]．算出法など詳細は成書に譲るが，このようにα係数は評価法内の一貫性をあらわす指標であり，検者間信頼性や検査—再検査信頼性と同じものを見ているものではないことに注意が必要である．α係数は評価尺度の中身をすべてまったく同じ項目にすると1になり，またまったく同じ項目を1つ加える（すなわち冗長にする）と必ず増加するという性質を有しているが，それにより必ずしも検者間信頼性が大きくなるわけではない．また，たとえばADLのような大きな概念の場合，単一の要素により構成されるとはかぎらず，たとえば，身辺活動や道具的活動，さらにはコミュニケーションなどの異なった複数の構成要素からなり立っている場合もある．このような場合は，α係数は低くなり，評価尺度が複数の構成要素を検査していることを示し，因子分析などを用いて大きな包括的概念を構成する要素的概念を検討することが必要となるが，その場合に必ずしも検査—再検査信頼性が小さくなるわけではないのである．このように，一口に信頼性といっても，異なった概念を意味していることがある．

## 2．妥当性

妥当性の評価法にはさまざまなものがあるが，大きく，基準関連妥当性criterion validity，構造概念妥当性construct validity，および内容妥当性content validityに分類されることが多い[6]．基準関連妥当性は，外的基準，すなわち測るべき概念の基準値（golden standard）との関連を見るものである．たとえば胸部X線写真にもとづき胸水貯留量を予測する場合に，予測値と実際の胸水を抜いてみた実測値との相関を見ることがこれにあたる．このように，計測時点でほぼ同時に相関を見る場合を同時的妥当性concurrent validityという．これに対し，時間的に後のことを予測するものを予測妥当性predictive validityという．たとえば閉所恐怖症の程度などを勘案したMRI検査可能尺度をつくった場合，この評価結果が，実際にMRI検査ができたかどうかをどの程度予測するかなどがこれにあたる．MRI検査ができるかどうかということがこの評価法の目標であり，実際にできたかどうかが明らかな基準値となるのである．このように基準関連妥当性は明らかな答えがある場合にのみ検討できる．

これに対し，構造概念妥当性は，想定された構造概念から理論的に予測される性質がその評価結果にみられるかどうかを検討するものであり，とくに明らかな外的基準がない場合に用いられる．たとえば脳血管障害により生じた記憶障害の評価法を作成する場合，他の同様な記憶の評価法の成績との相関は高く，言語や視覚認知などの記憶以外の認知機能評価法の成績との相関は低いことが示されなければならない．この場合前者を収束的妥当性convergent validity，後者を弁別的妥当性discriminant validityとよぶ．上述の同時的妥当性と収束的妥当性との相違は，相関を調べる対象が同時的妥当性では明らかな外的基準であるのに対し，収束的妥当性では外的基準ではなく，同じ構成概念を計測していると考えられている評価尺度であるという点である．同じ構成概念を計測していると考えられている評価尺度がない場合には，近い構成概念を計測している評価尺度との関連を見るときもある．たとえば記憶障害の評価法の場合は，全般的知能の評価尺度との相関を見ることがこれにあたる．この場合は，相関係数はそれほど高くはならない．

内容的妥当性は，個々の評価項目が，理論的に，または過去の文献を鑑みて，目的とする対象を測定するのに適切かどうかをあらわす．内容的妥当性の検討は統計解析ではなく，多くの専門家の判断を求めることによりおこなわれることが多い．これに対し，被検者が評価内容からこの検査が何を測ろうとしているかがわかることを，表面的妥当性face validityといい，内容的妥当性の一側面とされていることが多い．この表面的妥当性は被検者の反応に影響を与える．原則としては何を測ろうとしているのかが被検者にはっきりとわかるほうが適切な回答を得やすいが，たとえば性格の評価をおこなう場合などでは，まったく違うものを評価しているかのように被検者に思わせるほうが正しく評価できる場合もある．

前回に指摘したとおり，一般的には完璧な妥当性を有する測定方法は存在しないことが多い．たとえばADLのような曖昧な構成概念を完璧にとらえ網羅することができる評価尺度があるとは思われない．また外的基準があるものであれば高い妥当性を有するものが考えられるが，それらはたとえば脳梗塞巣の体積や脳血流量のような代用エンドポイントとしてしか用いられないものが多い．このため，エンドポイントとして用いられている評価尺度が，完全に妥当であるかどうかではなく，その研究目的のために十分な信頼性と妥当性が示されているか否かを批判的に吟味することが重要である．

## Ⅱ. 診断力の評価

診断力とはある検査がどの程度正確にある疾患や障害の有無を診断できるかという尺度である．これは診断という外的基準があり，それとの関連を評価するという点で上述の基準関連妥当性の一つであるといえる．診断力が高いとするためには，疾患や障害が実際にある患者をあると診断できるばかりではなく，実際にはない人をないと正しく診断できなければならない．ある疾患の検査をおこなった場合は，以下の4通りの結果が考えられる．

A：真陽性：実際に病気で検査結果も病気であった群
B：偽陽性：実際は正常なのに検査結果は病気であった群
C：偽陰性：実際は病気であったのに検査結果は正常であった群
D：真陰性：実際に正常で検査結果も正常であった群

この時，その検査の妥当性の指標として以下の尺度が用いられている．

感受性 sensitivity：実際に病気の人を病気とする割合 (A/(A＋C))
特異性 specificity：実際に正常な人を正常とする割合 (D/(B＋D))
陽性検出率 positive prediction rate：検査が病気のときに実際に病気の人の割合 (A/(A＋B))
陰性検出率 negative prediction rate：検査が正常のときに実際に正常な人の割合 (D/(C＋D))

この4つのなかで，一般的に用いられるのは感受性と特異性である．陽性および陰性検出率は感受性と特異性が同じでも病気の有病率により異なってくるため，あまりよい指標ではない．たとえば，感受性0.8，特異性0.8の脳梗塞の検査があり，若年者では脳梗塞の有病率が1％，高齢者では10％であったとすると，若年者10,000人に対し真の脳梗塞患者は100名，真の正常は9,900名，高齢者10,000人に対し真の脳梗塞患者は1,000名，真の正常は9,000名となる（表❷）．感受性と特異性がともに0.8であるから，若年者では真の患者のうち80名が検査陽性，正常者の9,900×(1－0.8)＝1,980名が検査陽性ということになり，陽性検出率は80/(80＋1,980)＝0.039となる．一方，高齢者では真の脳梗塞患者1,000名のうち800名が検査陽性となり，正常者9,000名のうち9,000×(1－0.8)＝1,800名が検査陽性ということになり，陽性検出率は800/(800＋1,800)＝0.31と大きくなることがわかる．

感受性と特異性はどちらかを高めると，どちらかが低くなるというトレードオフの関係にある．したがって，検査の診断力を比較する際には感受性あるいは特異性のどちらかの値だけではなく，両方の値を比較しておこなわなければならない．このように検査の診断力を比較する場合には受診者動作特性曲線 receiver operating characteristic curve（ROC曲線）が用いられる．これは縦軸に感受性，横軸に1-特異性すなわち偽陽性率を，プロットした曲線である．感受性も特異性も高い検査はROC曲線が左上方に凸になり，ROC曲線の下の面積（area under curve）が増加するため，これを用いて診断力を比較することができる[7][8]．

表❷ 若年者と高齢者の脳梗塞の有無と検査結果

A. 若年者

| | 脳梗塞あり | 脳梗塞なし | 合計 |
|---|---|---|---|
| 検査異常 | 80 | 1,980 | 2,060 |
| 検査正常 | 20 | 7,920 | 7,940 |
| 合計 | 100 | 9,900 | 10,000 |

B. 高齢者

| | 脳梗塞あり | 脳梗塞なし | 合計 |
|---|---|---|---|
| 検査異常 | 800 | 1,800 | 2,600 |
| 検査正常 | 200 | 7,200 | 7,400 |
| 合計 | 1,000 | 9,000 | 10,000 |

## Ⅲ. 各研究デザインにおける予測因子とエンドポイントとの関連の評価

### ◎1. コホート研究

コホート研究は，前述のようにある因子をもつ群ともたない群で，エンドポイントが生じた割合を比較し，その因子とエンドポイントとの関連を検証する研究デザインである．コホート研究で用いられる因子とエンドポイントとの関連の示標にはリスク比risk ratioあるいは相対危険度relative riskと，リスク差risk differenceあるいは寄与リスクattributable riskがある．リスク比は，当該因子をもつ群のエンドポイント発生率を，当該因子をもたない群での発生率で割った値で，当該因子があると何倍エンドポイントが起きやすくなるかをあらわす指標であり，その因子がエンドポイントとまったく無関係であれば1となる．リスク比が1より大きいほど，その因子はそのエンドポイントの発生に強く促進的にかかわり，反対に1より小さくなり0に近づくほど，その因子はそのエンドポイントの発生に強く防御的にかかわることをあらわしている．リスク差は当該因子をもつ群のエンドポイント発生率から，当該因子をもたない群での発生率を引いた値で，当該因子があることにより実際にどれだけエンドポイント発生率が増えるかを示している．たとえば，喫煙という因子の有無により脳梗塞の発症というエンドポイントの発生を観察したコホート研究で，それぞれ10,000人の観察で，喫煙群で100人，非喫煙群で10人の脳梗塞の発生があった場合も，それぞれ1,000人と100人の発生があった場合もリスク比は10であるが，リスク差は前者が90/10,000，後者は900/10,000となり，喫煙による脳梗塞の発生数の増加は，後者のほうが明らかに大きくなる．

### ◎2. 症例対照研究

症例対照研究では，ある疾病を有している患者と有していない対照者のあいだで，ある因子を有していたものの割合を比較検討する研究デザインである．症例対照研究では直接発生率を見ることができず，リスク比やリスク差をその指標として用いることはできない．このため，症例対照研究で用いられる因子とエンドポイントとの関連の示標にはオッズ比odds ratioが用いられる．これは疾病群における当該因子をもつ人数をもたない人数で割った値（オッズ）を，対照群における同じオッズでさらに割った値である．オッズ比は当該疾患の有病率が小さい場合はリスク比に近似でき，リスク比の場合と同じくオッズ比が1から遠ざかるほどその因子とそのエンドポイントとの関連が強いことを意味している．リスク比もオッズ比も95％信頼区間を計算することが可能であり，95％信頼区間が1をはさんでいない場合には，その因子は有意な危険あるいは防御因子であると考えられる．

### ◎3. 無作為化対照試験

無作為化対照試験などの実験疫学研究は，介入interventionにより，因子を操作することによって，エンドポイントの発生がどれくらい低減するかを検討するためにおこなわれる．実験疫学研究は，基本的には因子への介入の有無で分けた2群によるコホート研究であり，コホート研究で用いられるリスク比，リスク差をその示標として用いることができる．その他にも実験疫学研究に特有な評価法として以下のものがあげられる．

**相対リスク減少率relative risk reduction（RRR）**

相対改善度relative improvement ratioともいう．介入群と対照群におけるエンドポイント発生率の差を対照群におけるエンドポイント発生率で割った値であり，介入による，エンドポイントの発生の減少率を示したものである．

**絶対リスク減少率absolute risk reduction（ARR）**

これは介入群におけるエンドポイント発生率と対照群におけるエンドポイント発生率との差である．この絶対リスク減少率の逆数を，number needed to treat（NNT）といい，エンドポイントの発生が一人でみられなくなるために介入することが必要な人数である．NNTが小さいほど効果的な介入法であることを示している．

## おわりに

今回までは，基礎的な臨床疫学および生物統計学を紹介してきた．臨床疫学の知識はややもすれば難解であるが，前述したように論文の批判的吟味をするためにはどうしても必要な知識であり，しばらくのあいだおつきあいをいただいた．次回からは，論文の読み取り方について具体的な例をあげながら説明していきたいと思う．

●文　献●

1) Cohen J：A coefficient of agreement for nominal scales. *Educ Psychol Meas* **20**：37-46, 1960
2) Lyden PD *et al*：A critical appraisal of stroke evaluation and rating scales. *Stroke* **22**：1345-1352, 1991
3) Fleiss JL：Measuring nominal scale agreement among many raters. *Psychol Bull* **76**：378-382, 1971
4) Cohen J：Weighted Kappa：Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol Bull* **70**：213-220, 1968
5) Dunn G：Design and analysis of reliability studies. Edward Arnold, London, 1989
6) Franzen MD：Reliability and validity in neuropsychological assessment. Plenum Press, New York, 1989
7) Hanley JA *et al*：The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143**：29-36, 1982
8) Hanley JA *et al*：A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology* **148**：839-843, 1983

原著

# 軽症アルツハイマー病患者における リバーミード行動記憶検査の有用性

松田明美[1) ] 数井裕光[2)] 博野信次[2)]
森 悦朗[2)]

**要旨** 〈目的〉日本版リバーミード行動記憶検査(RBMT)の軽度アルツハイマー病(AD)における有用性の検討。〈方法〉軽症AD患者100名に，RBMTを施行し，RBMTとWechsler Memory Scale-Revisedなどの他の記憶検査や，記憶障害による日常生活上の障害を評価する生活健忘チェックリストとの相関を検討した。また，年齢，性別，教育歴をマッチさせたAD患者と健常対照者(NC)46組のRBMTの結果を比較検討した。〈結果〉RBMTは既存の記憶検査や日常生活上の障害と有意に相関し，特に日常生活上の障害と最も高い相関を示していたのはRBMTであった。AD群とNC群の比較では，RBMTの成績はAD群で有意に低く，またAD患者と健常者をそれぞれ95%以上，正しく分類することができた。〈結論〉RBMTはAD患者の日常記憶の評価およびAD患者と健常者との鑑別診断に有用である。

**Key words**：the Japanese version of Rivermead Behavioural Memory Test, Alzheimer's disease, everyday memory

## はじめに

アルツハイマー病(AD)の多くは，記憶障害から発症し，見当識障害，言語障害，構成障害，失行，失認などがそれに引き続いて生じる[9]。しかも一般的に記憶障害はこれらの障害の中でも中核であり，日常生活障害は記憶障害によって起こることが多い。そのためAD患者の日常生活における記憶機能，すなわち日常記憶を評価することは重要である。AD患者の日常記憶の障害を把握することにより，日常生活上の問題を予測することができ，リハビリテーションスタッフおよび医療関係者は，家族や介護者に対して，より適切な指導ができると考えられる。

日常記憶とは，実際の日常生活場面で必要とされる記憶のことで，その中には建物などの場所の記憶，顔や名前の記憶，会話の記憶，展望記憶，自伝的記憶などが含まれる[4]。日常記憶の障害を評価する検査として，Rivermead Behavioural Memory Test(RBMT)[12]が開発され，欧米では広く使用されている。最近，綿森らは，原版著者の許可を得て本邦の実状に合うよう写真刺激，物語などの一部を改変しつつ，概ね原版を忠実に翻訳し日本版RBMTを作成し，われわれと共同してその標準化を行った[4]。RBMTは，頭部外傷や脳血管障害などで記憶障害を有する脳損傷患者において有用であることは，これまで報告されている[1,5-7]が，記憶障害だけでなく，その他の領域の認知機能も障害されるAD患者における有用性を検討した報告はなかった。今回，われわれは軽度のAD患者を対象に日本版RBMTを施行し，この検査のAD患者の日常記憶の評価における有用性を検討したので報告する。

## I. 対象と方法

〈RBMT〉

RBMTは英国オックスフォードのリバーミード・リハビリテーションセンターで，日常記憶の障害を発見し，また治療による変化を観察するために開発された記憶検査バッテリーである。日常記憶の障害には様々な認知機能障害が影響していると考えられるが，

1) 兵庫県立高齢者脳機能研究センターリハビリテーション科 2) 同 臨床研究科(2002年3月15日受稿)
〔連絡先〕松田明美：兵庫県立高齢者脳機能研究センターリハビリテーション科(〒670-0981 兵庫県姫路市西庄甲520)

RBMTは認知機能障害の影響も含めて，記憶障害が日常生活にどのように影響しているかを調べることを目的としている。

またRBMTでは，日常記憶をより自然な形で検査できるように，日常生活の中で脳損傷患者の健忘症状が明らかになるような場面を検査室内で可能な限り再現するような工夫がなされている。RBMTは9個(細かく分類すると12個)の下位検査から構成されているが，その下位検査の選択には，これまでの記憶研究による分類が織り込まれ，多様な記憶形態を測定できるようになっている。すなわち，記憶の素材として言語的項目と視覚的・空間的項目が設定され，時間的分類からは展望記憶と回顧的記憶(過去の記憶)の課題が含まれている。さらに回顧的記憶の課題のいくつかでは直後に加え，一定時間後に遅延検査するようになっている。

RBMTには日常記憶の検査であるという特徴の他に2つの優れた点がある。一つは同等の難易度であることが確認されている並行バッテリーが4つ用意されていることである。このため繰り返し施行による練習効果の影響を排除して日常記憶の様態を縦断的に評価することができる。すなわち薬物治療をはじめとする治療的介入による記憶機能の改善の評価，あるいは逆に記憶障害の悪化の評価などに用いることができる。もう一つはこれまでの包括的な記憶検査よりも短時間で施行でき，かつ指示・手順も簡単なので，比較的重症の症例にも施行可能であることである。

RBMTに含まれる下位検査は以下の通りである：(1)1枚の顔写真を被験者に見せて，その人の姓・名を記憶させ，遅延再生させる課題，(2)被験者の持ち物を借りて隠し，検査の終了時に被験者がその持ち物の返却を要求する約束の記憶と，その持ち物を隠した場所の記憶，(3)20分後に鳴るようにアラームをセットし，アラームが鳴ったら，決められた質問をするという約束の記憶，(4)絵カードの遅延再認課題，(5)短い物語の直後再生と遅延再生，(6)顔写真の遅延再認課題，(7)検者が部屋の中で一定の道順を辿ってみせ，それを直後と遅延を置いた後に被験者に辿らせる課題，(8)道順の中である用件を行わせる課題，(9)見当識の課題。

RBMTの採点は個々の下位検査ごとに行われる。それぞれの素点を，規定の変換法に従い標準プロフィール点とスクリーニング点に換算する。標準プロフィール点は，各下位検査ごとの基準に従って0～2点の3段階に換算され，スクリーニング点は，満点なら1点，それ以外は0点という基準で換算される。標準プロフィール点は下位検査間の成績を直接比較できるようにするために，それぞれの下位検査の難易度を考慮し換算された得点である。生活健忘の全般的な指標としては，標準プロフィール点合計とスクリーニング点合計が用いられ，それぞれの満点は24点および12点である。

〈対象〉

対象は兵庫県立高齢者脳機能研究センター附属病院に精査のために入院した患者のうち，National Institute of Neurological and Communicative Disorders and Stroke/Alzheimer's Disease and Related Disorders Association の probable Alzheimer's disease の診断基準を満たし，Clinical Dementia Rating(CDR)による痴呆重症度がごく軽度(0.5)あるいは軽度(1)のAD患者100例である。性別は男性27例，女性73例，平均年齢は71.6±7.8(SD)歳，平均教育歴は9.6±2.4年，Mini-Mental State Examination(MMSE)の平均点は22.7±3.3点，CDRは0.5が11例，1が89例であった。CDRの記憶下位項目の得点は0.5が3例，1が46例，2が40例，3が11例であった。MRI上，軽度の白質病変を除く局所性病変を認める症例や，認知機能障害が生じる可能性のあるAD以外の合併症を有する症例は除外した。また全例でpositron emission tomography(PET)あるいはsingle photon emission computed tomography(SPECT)で，両側の側頭葉内側部か両側頭頂葉の血流あるいは代謝低下を確認している。

次に，RBMTの検査結果を知らない神経内科医1名が，AD患者100例とRBMTの標準化研究にボランティアとして参加した健常被験者274例の中から，年齢，性別，教育歴を一致させたADと健常対照者(NC)の46組を選び出した。選択基準は性別，教育年数には全く差がなく，年齢の差は±3歳とした。その46組の性別は男性17例，女性29例ずつ，平均教育歴は10.3±1.9年であった。AD群の平均年齢は69.0±8.1歳，MMSEの平均点は22.7±3.5点，CDRは0.5が7例，1が39例であった。NC群の平均年齢は68.6±8.0歳で，MMSEの平均点は27.7±1.9点であった。

〈方法〉

RBMTの4つの並行バッテリーの中のA版を全例に施行した。さらにRBMTの記憶検査としての妥当性を検討するために，Wechsler Memory Scale-Revised(WMS-R)[10]とAlzheimer's Disease Assessment Scale(ADAS)[3]を施行した。そしてWMS-Rの言語性記憶検査加重合計得点，視覚性記憶検査加重合計得点，全般性記憶検査加重合計得点，遅延再生検査加重合計得点の成績およびADASの単語再生課題の合計正再

Table 1 Spearman rank correlation coefficients between scores of the Japanese version of Rivermead Behavioural Memory Test(RBMT) and scores of the Wechsler Memory Scale-Revised(WMS-R) and of the Alzheimer's Disease Assessment Scale-Cognitive part(ADAS-J Cog).

| RBMT | total profile score | total screening score |
|---|---|---|
| WMS-R | | |
| verbal weighted sum score | 0.47*** | 0.34*** |
| visual weighted sum score | 0.57*** | 0.42*** |
| general weighted sum score | 0.60*** | 0.44*** |
| delayed weighted sum score | 0.62*** | 0.53*** |
| ADAS-J Cog | | |
| word recall subtest score | 0.33*** | 0.24* |

* $p<0.05$, *** $p<0.001$

Table 2 Spearman rank correlation coefficients between scores of the Memory Checklist(CL) and memory subscale of the Clinical Dementia Rating(CDR) and scores of the Japanese version of Rivermead Behavioural Memory Test(RBMT), the Wechsler Memory Scale-Revised(WMS-R), and the Alzheimer's Disease Assessment Scale-Cognitive part(ADAS-J Cog).

| | CL | CDR memory subscale |
|---|---|---|
| RBMT | | |
| total profile score | −0.25* | −0.21* |
| total screening score | −0.32** | −0.31** |
| WMS-R | | |
| verbal weighted sum score | −0.13 | −0.05 |
| visual weighted sum score | −0.19 | −0.08 |
| general weighted sum score | −0.18 | −0.08 |
| delayed weighted sum score | −0.30** | −0.20* |
| ADAS-J Cog | | |
| word recall subtest score | −0.16 | −0.11 |

*$p<0.05$, **$p<0.01$

生数とRBMTの各得点との相関を検討した。

また，記憶障害によって生じる患者の日常生活場面での実際の問題の有無およびその頻度を，生活健忘チェックリスト(CL)[4]を用いて，AD患者の介護者から調査した。CLはWilsonらのMemory Checklist[11]を著者の許可を得て，われわれと綿森らが本邦の実状に合うように一部改変を加えながら原版に忠実に翻訳，作成したものである。その中では，日常記憶の障害のために実生活で起こりうる問題あるいはそのような場面が13項目設定され，それぞれの項目について最近1カ月間の頻度が0(全くない)～3(常にそうである)の4段階で評価され，それらの合計が得点となる。高得点になるほど日常生活上の問題が強いということになり，満点は39点である。このCL得点，および患者の日常生活の観察による日常記憶能力評価であるCDR記憶下位項目得点とRBMTの各合計点，WMS-RおよびADASの各得点との相関を検討し，各記憶検査の日常記憶の評価としての妥当性を検討した。

次に，RBMTがAD患者とNCとの鑑別に有用かどうかを検討するために，年齢，性別，教育歴を一致させたAD群とNC群の，RBMTのA版の標準プロフィール点合計とスクリーニング点合計をMann-WhitneyのU検定で比較するとともに，それぞれの得点の鑑別診断力を検定した。

## II. 結　果

RBMTの日常生活記憶検査としての妥当性についての結果を示す。AD群において，RBMTの標準プロフィール点合計，スクリーニング点合計は，それぞれWMS-Rの言語性記憶得点，視覚性記憶得点，全般性記憶得点，遅延再生得点，およびADASの単語再生課題の合計正再生数全てと有意な相関を示した(Table

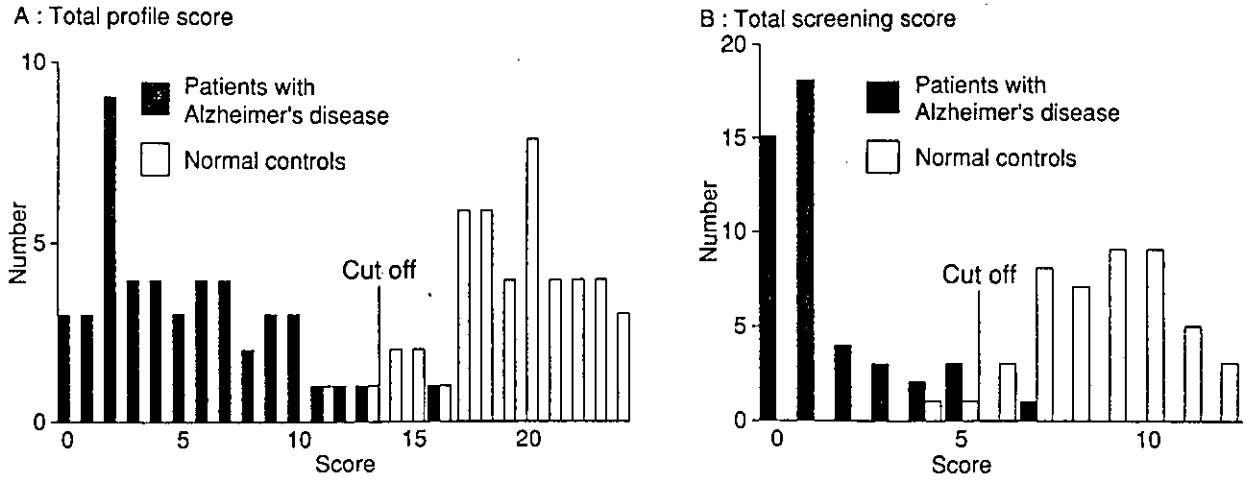



**Fig. 1** The result of the Japanese version of Rivermead Behavioural Memory Test in the patients with Alzheimer's disease and normal controls.

1)。

また，CL得点およびCDRの記憶下位項目得点はRBMTの標準プロフィール点合計，スクリーニング点合計とWMS-Rの遅延再生得点と有意な相関を示し，特にRBMTのスクリーニング得点と高い相関を示していた(Table 2)。

次に，RBMTの鑑別能についての結果を示す。RBMTの標準プロフィール点合計は，AD群が5.2±3.8点で，NC群が19.1±3.0点(U=9.5, $p<0.001$)，スクリーニング点合計は，AD群が1.4±1.7点で，NC群が8.7±1.9点(U=14.5, $p<0.001$)であり，両得点ともAD群はNC群と比較し有意に低かった。

また標準プロフィール点合計のカットオフ値を14/13とすると，AD患者の98.8%，NCの95.7%を正しく分類することができた。スクリーニング点合計では，カットオフ値を6/5点とすると，AD患者の97.8%，NCの95.7%を正しく分類することができた(Fig. 1)。

## III. 考 察

今回の検討から，AD群のRBMTの標準プロフィール点合計，スクリーニング点合計は，既存の記憶検査であるWMS-RおよびADASの下位検査の成績と有意な相関を示し，RBMTはADの記憶機能の評価として妥当であることが示された。さらにRBMTの標準プロフィール点合計，スクリーニング点合計はCLで評価された日常生活場面での問題やCDRの記憶下位項目で評価された患者の日常生活の観察による日常記憶能力評価との間に有意な相関を示し，さらに本検査が有する特性から，RBMTは日常記憶の評価に有用であるといえる。またAD群ではNC群に比べて，RBMTの標準プロフィール点合計，スクリーニング点合計が有意に低く，RBMTの標準プロフィール点合計，スクリーニング点合計によりそれぞれ95%以上の正確さでAD患者とNCとを鑑別することができた。このことから，RBMTはAD患者とNCとの鑑別診断にも有用であるといえる。

AD患者の記憶機能の評価法には，様々な種類があり，評価の形式も異なっている。記憶検査には，患者に対する質問形式で評価するものと，患者をよく知る介護者などから情報を得て，日常生活の行動観察によって評価するものとに分けられる。患者に対する質問形式で評価するものには，MMSEやADAS，WMS-Rの他に，長谷川式簡易知能評価スケール，N式精神機能検査などがある。これらの検査は，見当識や言語的・視覚的記憶などの記憶機能の評価だけでなく，構成能力や注意機能などの他の認知機能の評価も含まれており，痴呆症状の重症度を評価するのに有用であるとされている。一方介護者などの行動観察に基づき，AD患者の記憶機能を評価するものには，CLやCDRの記憶下位尺度，GBSスケール，N式老年者用精神状態尺度，Subjective Memory Questionnaire[8]，Everyday Memory Questionnaire[2,8]などがあり，患者をよく知る介護者などから情報を得て，AD患者の日常生活動作，および記憶機能を把握することができる。質問形式による評価および介護者などの行動観察に基づく評価は，それぞれ長所・短所を持っているため，これらを組み合わせることによって，患者を評価することが重要である。

これに対して，RBMTはこれまでの記憶検査と異なる優れた特徴が3つある。1つ目の利点は，RBMT

では検査室で日常的な場面を設定し，検者が患者を直接観察することで評価できることである。これによって，患者の記憶障害が日常生活にどのように影響しているかを，より具体的に把握することができる。これまで，検者が直接観察することによって，記憶機能を評価する方法はなかった。RBMTの2つ目の利点は，同等の難易度である並行バッテリーが4つ用意されているため，繰り返しによる練習効果の影響を排除することができ，効果判定などに役立つことである。3つ目の利点は，RBMTはADASやWMS-Rなどの既存の検査よりも，短時間で施行可能ということである。AD患者のADAS，WMS-Rの所要時間は，それぞれ約40分，60分かかるといわれているのに対し，RBMTは約30分で施行可能である。今回のAD患者100名における平均所要時間は27.3分であった。このようにRBMTは比較的短時間に日常記憶能力を把握でき，AD患者と健常人との鑑別も行えることから，ADの診察上，非常に有用な検査であると考えられる。

今回の検討でAD患者におけるRBMTの有用性が明らかにできた。RBMTはAD患者の日常記憶を評価でき，患者の日常生活上での問題を予測し家族や介護者に対する介入法の指導に役立つものと期待できる。

## 文　献


1) Clare L, Wilson BA, Carter G, Breen K, Gosses A, Hodges JR : Intervening with everyday memory problems in dementia of Alzheimer type : An errorless learning approach. J Clin Exp Neuropsychol 22 : 132–146, 2000
2) Cornish I M : Factor structure of the Everyday Memory Questionnaire. Br J Psychol 91 : 427–438, 2000
3) 本間　昭, 福沢一吉, 塚田良雄, 石井徹郎, 長谷川和夫, Mohs RC : Alzheimer's Disease Assessment Scale (ADAS)日本版の作成. 老年精神医学3 : 647–655, 1992
4) 数井裕光, 綿森淑子, 本多留美, 時政昭次, 博野信次, 森悦朗 : 日本版リバーミード行動記憶検査(RBMT)の有用性の検討. 神経進歩46 : 307–318, 2002
5) Kotler–Cope S, Camp CJ : Anosognosia in Alzheimer Disease. Alzheimer Dis Assoc Disord 9 : 52–56, 1995
6) Ownsworth T, McFarland K : Memory remediation in longterm acquired brain injury : Two approaches in diary training. Brain Inj 13 : 605–626, 1999
7) Perez M, Godoy J : Comparison between a "Traditional" memory test and a "Behavioral" memory battery in Spanish patients. J Clin Exp Neuropsychol 20 : 496–502, 1998
8) Schwartz A F : Assessment of the everyday memory after severe head injury. Cortex 25 : 665–671, 1989
9) Small GW, Rabins PV, Barry PP, Buckholtz NS, DeKosky ST, Ferris SH, Finkel SI, Gwyther LP, Khachaturian ZS, Lebowitz BD, McRae TD, Morris JC, Oakley F, Schneider LS, Streim JE, Sunderland T, Teri LA, Tune LE : Diagnosis and treatment of Alzheimer disease and related disorders : Consensus statement of the American Association for Geriatric Psychiatry, the Alzheimer's Association, and the American Geriatrics Society. JAMA 278 : 1363–1371, 1997
10) Wechsler D : A standardized memory scale for clinical use. J Psychol 19 : 87–95, 1945
11) Wilson B, Cockburn J, Baddeley A, Hiorns R : The development and validation of a test battery for detecting and monitoring everyday memory problems. J Clin Exp Neuropsychol 11 : 855–870, 1989
12) Wilson B, Cockburn J, Baddeley A : The Rivermead Behavioural Memory Test, Thames Valley Test Company, England, 1991


## Abstract

Validity of the Japanese Version of Rivermead Behavioural Memory Test for Evaluation of Everyday Memory Function in Patients with Mild Alzheimer's Disease


by

Akemi Matsuda[1], Hiroaki Kazui[2], Nobutsugu Hirono[2], Etsuro Mori[2]

from

Divisions of Neurorehabilitation[1] and Clinical Neurosciences[2], Hyogo Institute for Aging Brain and Cognitive Disorders, 520 Saisho-Ko, Himeji, 670–0981 Hyogo, Japan



**Objective** : To validate the Japanese version of Rivermead Behavioural Memory Test(RBMT) in evaluating everyday memory function in patients with mild Alzheimer's disease(AD). **Subjects and Methods** : Subjects were 100 patients with probable AD of very mild or mild stages of dementia as measured by the Clinical Dementia Rating(CDR). Scores of the Japanese version of RBMT were correlated with scores of the Alzheimer's Disease Assessment Scale–Cognitive part (ADAS–J Cog) and of the Wechsler Memory Scale–Revised(WMS–R). The everyday disability caused by the impairment of memory function was also assessed with the Memory Checklist(CL), and was correlated with the RBMT scores. In addition, the diagnostic value of the RBMT was examined in 46 pairs of AD patients and healthy subjects ; the patients were chosen from the present subjects and the healthy subjects from the participants in the previous RBMT standardization study as one–by–one matched for age, sex, and educational level. **Results** : Both of the total screening score and profile score of RBMT were significantly correlated with all of the weighted sum scores of ver-

bal, visual, general, and delayed memory tests of the WMS-R and the word recall subtest score of the ADAS-J Cog. Both of the RBMT scores were also highly correlated with the CL score and the memory subscale of the CDR. In the analysis of the diagnostic accuracy, both of the RBMT scores correctly classified 98% of AD patients and 96% of normal volunteers by setting the cut-off scores of 6/5 for the total screening score and of 14/13 for the total profile score. **Conclusions**: The RBMT is a useful tool in evaluating everyday memory function in patients with mild AD. This test also accurately differentiates AD patients from healthy individuals.

(Received: March 15, 2002)

特集／ぼけ・痴呆－診療と予防

# アルツハイマー型痴呆の初期診断に必要な高次機能検査・画像検査

橋本　衛　森　悦朗

## はじめに

物忘れを主訴として来院した患者が，初期のアルツハイマー型痴呆なのかそれとも生理的な加齢の範囲内なのか，診断に難渋するケースがしばしばある。高齢化社会が進行するにつれこのようなケースは増加し，一方新しい抗痴呆薬の開発とともにアルツハイマー型痴呆の早期診断がますます必要とされて来る。今回の特集では，アルツハイマー型痴呆の早期診断に重点を置き，われわれのセンターで使用している高次機能検査と画像検査を中心に紹介する。

## Ⅰ．高次機能検査

### 1．スクリーニング検査

痴呆が疑われる患者を診察する場合，まず全般的な認知機能を評価する目的でスクリーニング検査を実施する。これらは診察場面で簡単に行える検査であり，スクリーニング検査で異常があれば更に詳細な心理検査が必要となる。

#### 1）Mini-Mental State Examination (MMSE)[1]（図1）

世界的に最も一般的に用いられている痴呆の簡易検査である。見当識，記銘，注意と計算，再生，言語，構成の6項目から構成され満点は30点である。施行時間は約20分で，23～24点以下の場合痴呆が疑われる。記憶機能検査に重点が置かれているため，アルツハイマー型痴呆のスクリーニング検査として適している。またアルツハイマー型痴呆で障害されやすい言語機能や視覚構成機能を検査する項目も含まれているため，典型的なアルツハイマー型痴呆患者であればMMSEを行うだけでおおよその見当はつく。兵庫脳研方式（MMSE）では，認知障害のある患者の83.8％が23点以下であった。

兵庫県立高齢者脳機能研究センター臨床研究科

#### 2）改訂長谷川式簡易知能評価スケール (HSD-R)[2]

わが国で最も用いられている痴呆の簡易検査（満点は30点）である。20点以下は痴呆が疑われる。記憶，見当識，注意，計算に加えて，野菜の名前を列挙する実行機能検査が含まれている。初期のアルツハイマー型痴呆で障害される記憶機能と実行機能を評価する項目に重点が置かれているため，アルツハイマー型痴呆のスクリーニング検査として適している。言語機能や構成機能を評価する項目は含まれておらず，認知機能障害の全般的評価法として用いることはできない。

ごく初期のアルツハイマー型痴呆であればしばしばスクリーニング検査の結果が正常範囲内を示す。仮に正常範囲内の得点であっても，病歴や問診で痴呆が強く疑われる場合はさらに詳細な心理検査を行うべきである。

### 2．複雑な検査

#### 1）Wechsler Memory Scale-Revised (WMS-R)[3]

WMS-Rは，欧米で広く使われている記憶検査である。ごく最近日本語版が出版された。MMSEやADASで検査されるような言語性の記憶に加えて，視覚性の記憶機能を検査できる利点がある。WMS-Rでは，見当識，言語性記憶，視覚性記憶，注意機能および遅延後の記憶能力を調べる13の下位検査からなり，各機能のindexを算出し，数値で比較できるようになっている。われわれの検討では，WMS-R（特に遅延記憶）は健常高齢者群とCDR 0.5のごく軽度アルツハイマー型痴呆群とを感受性，特異性いずれも90％以上の鑑別能力を示し[4]，アルツハイマー型痴呆の早期診断に極めて鋭敏な検査である。

#### 2）Wechsler adult intelligence scale-revised (日本版 WAIS-R)[5]

WAIS-Rは現在日本で広く使用されている知能

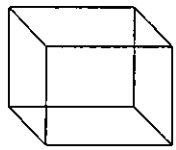
| | |
|---|---|
| 1. 時の見当識（5点） | 8. 3段階の命令（3点） |
| 「今年は何年ですか」「今の季節は何ですか」「今日は何月ですか」「今日は何日ですか」「今は何時ですか」　各1点 | 机上に大小2枚の紙を置き、次の文を言い、その通りにやってもらう。「小さい方の紙を取り、半分に折って大きい紙の下に入れる」「小さい方の紙を取る」で1点「半分に折る」で1点、「大きい紙の下に入れる」で1点 |
| 2. 場所の見当識（5点） | 9. 読んで従う（1点） |
| 「ここは何県ですか」「ここは何市ですか」「ここは何病院ですか」「ここは何階ですか」「ここは何号室（何科）ですか」 | 次の文を読んで、その指示に従ってください。「目を閉じてください」 |
| 3. 記銘（3点） | 10. 文を書く（1点） |
| 相互に無関係な3つの語（EX. 犬、桜、電車）を検者が1秒間に1語ずつ言う。3つ言った後で何であったかを尋ねる。正答1個につき1点を与える。3個全て言えるまで繰り返し、繰り返し回数を記録する。 | なにかを書いてください。 |
| 4. 注意（5点） | 11. 図形の模写（1点） |
| 100から順に7をひいた答えを言ってもらう（5回まで）。正答1つにつき1点。途中の式は与えない（EX.93ひく7は、とは尋ねないこと）。 |  |
| 5. 再生（3点） | |
| 記銘から5分後に、先に繰り返した3つの語を尋ねる。正答1つにつき1点。 | |
| 6. 呼称（2点） | |
| （時計を見せながら）これは何ですか。（鉛筆を見せながら）これは何ですか。 | |
| 7. 復唱（1点） | |
| 次の文を言い、繰り返してもらう。「ちりもつもればやまとなる」 | |

図　1　兵庫県立高齢者脳機能研究センター版 Mini-Mental State Examination (MMSE)

検査である。WAIS-R は11種の下位検査からなり，全検査の知能指数（IQ）に加えて，言語性 IQ と動作性 IQ がそれぞれ算出できるのが特徴である。また，11種の下位検査の成績に基づいたプロフィールを描き，知的機能の各項における強弱や相互の関連性を示すことが可能である。知能検査のプロフィールを検討することで，患者の病前の能力から低下した知的機能を明らかにしていくが，何点以下が痴呆であるといった線引きはない。また，初期のアルツハイマー型痴呆で記憶のみが強く障害されているケースでは正常値を示すことが多く，初期のアルツハイマー型痴呆と正常者との鑑別よりもむしろ，アルツハイマー型痴呆とその他の痴呆の鑑別により有用な検査である。

**3）Alzheimer's Disease Assessment Scale-cognitive subscale 日本語版（ADAS-J cog.）[6][7]**

アルツハイマー型痴呆を対象とした抗痴呆薬の臨床試験における認知機能検査として国際的に用いられている。記憶，言語，行為・構成の3領域に関する，計11の下位検査項目から構成され，特に記憶の評価に重点がおかれている。通常は認知機能障害の縦断的評価法として利用されているが，障害のプロフィールをみることによって，痴呆の鑑別診断に用いることも可能である。われわれのセンターでの検討では，認知障害なしの上限を9点，認知障害ありの下限を10点とすると，アルツハイマー型痴呆患者と正常高齢者の鑑別において，高い感度（98.1%）と特異性（95.1%）が得られた[8]。

**3．発症を予測する検査**

将来アルツハイマー型痴呆を発症するかどうかを痴呆症状が明らかとなる前の段階で予測する試みが近年盛んに行われている。Chen らは痴呆のない高齢被験者を対象として縦断的に多数の認知機能検査を行い，どの検査の低下が最もアルツハイマー型痴呆の発症を強く予測するかを検討した[9]。その結果，単語再生課題（遅延再生がより有効）と Trail-Making Test の成績がアルツハイマー型痴呆の発症をより強く予測することを示した。Trail-Making Test は実行機能を評価する検査と考えられており，彼らの結果はアルツハイマー型痴呆の初発症状が記憶の遅延再生と実行機能の障害である可能性を示唆している。

Trail-Making Test は紙面上に配置された数字を線で順番に結んでいく検査である[10]（図2）。Part A と Part B の二種類があり，課題は Part B の方が難しく，Chen らの検討では Part B の方がより強くアルツハイマー型痴呆の発症を予測する結果となっている。所要時間は10分程度と短く，診察場面でのスクリーニング検査として手軽に利用できる。

単語再生課題については，ADAS の一項目として実施されておりこれを利用することが手軽な方法

であろう。ただしADASの場合即時再生のみで遅延再生は含まれていない。われわれのセンターでは通常のADASに単語再生課題の遅延再生と単語再認課題の遅延再認を加えることにより，ごく初期のアルツハイマー型痴呆患者を見逃さない工夫をしている。

**4．心理検査実施上の注意点**

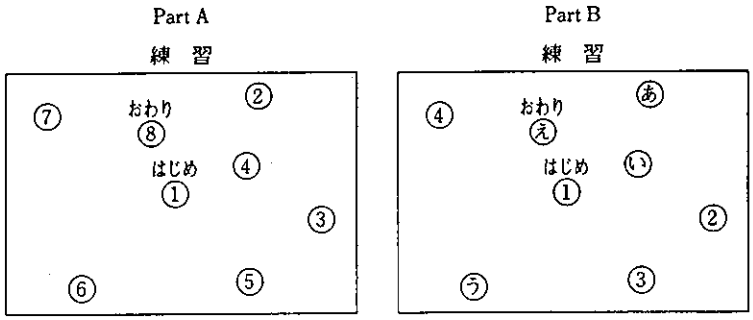
心理検査は痴呆患者にとって最も苦手なことを強いるため，しばしば緊張や不安，防衛を引き起こす。検査者は検査の目的を患者に説明し，患者とのコミュニケーションをしっかり取ることが重要である。高齢者では視力や聴力の低下がよくあり，きめ細かい配慮や検査方法の工夫が必要である。検査の信頼性を高め，得られた成績に正しい解釈を与えるためには，熟練した検査者が心理検査を行うことが求められる。また，心理検査の結果は病前の知的レベルに影響されるため，テストの結果を評価する際には教育歴や職歴などを考慮しなければならない。

## Ⅱ．画　像　検　査

**1．形態画像検査**

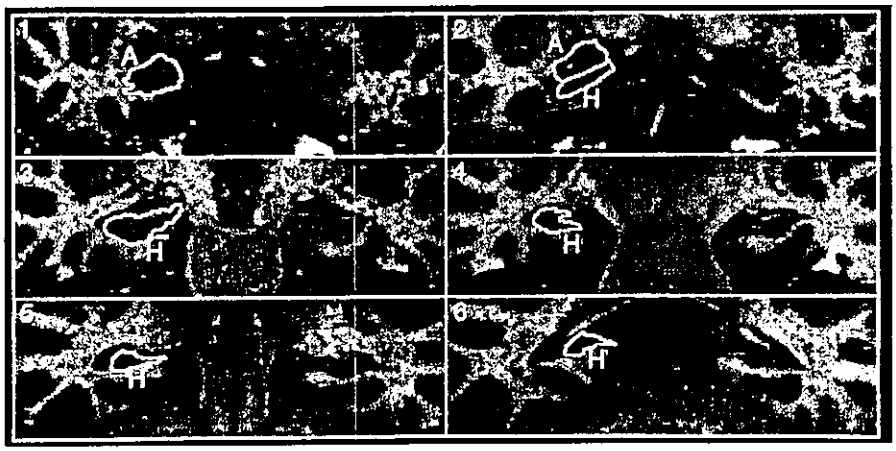
アルツハイマー型痴呆はこれまで画像検査上は特異的な変化が認められず，とりわけ病初期に異常を指摘することは困難と考えられ，主に臨床症状によって診断されるものであった。しかし，CTやMRIの開発とともに，梗塞や出血などの血管性病変を除外し，アルツハイマー型痴呆に特異的な脳萎縮を同定することが可能となり，アルツハイマー型痴呆の診断にこれらの形態画像検査が積極的に用いられるようになってきた。

アルツハイマー型痴呆の肉眼的な神経病理学的変化としては脳萎縮が代表される。この萎縮を定量的に評価してアルツハイマー型痴呆の診断に用いようとする試みはCTの時代から行われてきた。とりわけアルツハイマー型痴呆では病初期より海馬の萎縮が認められることが特徴的であることから，海馬の




Part Aは数字を順番に，Part Bは数字と仮名を順に（例：1-あ-2-い-）結ぶ。課題の達成時間を評価する。

図　2　Trail Making Testの練習課題




71歳の軽症アルツハイマー病女性。海馬は軽度に萎縮している。1～6は計測の最前部から最後部を示す。海馬は海馬体に海馬台を含む。
A：扁桃体，H：海馬

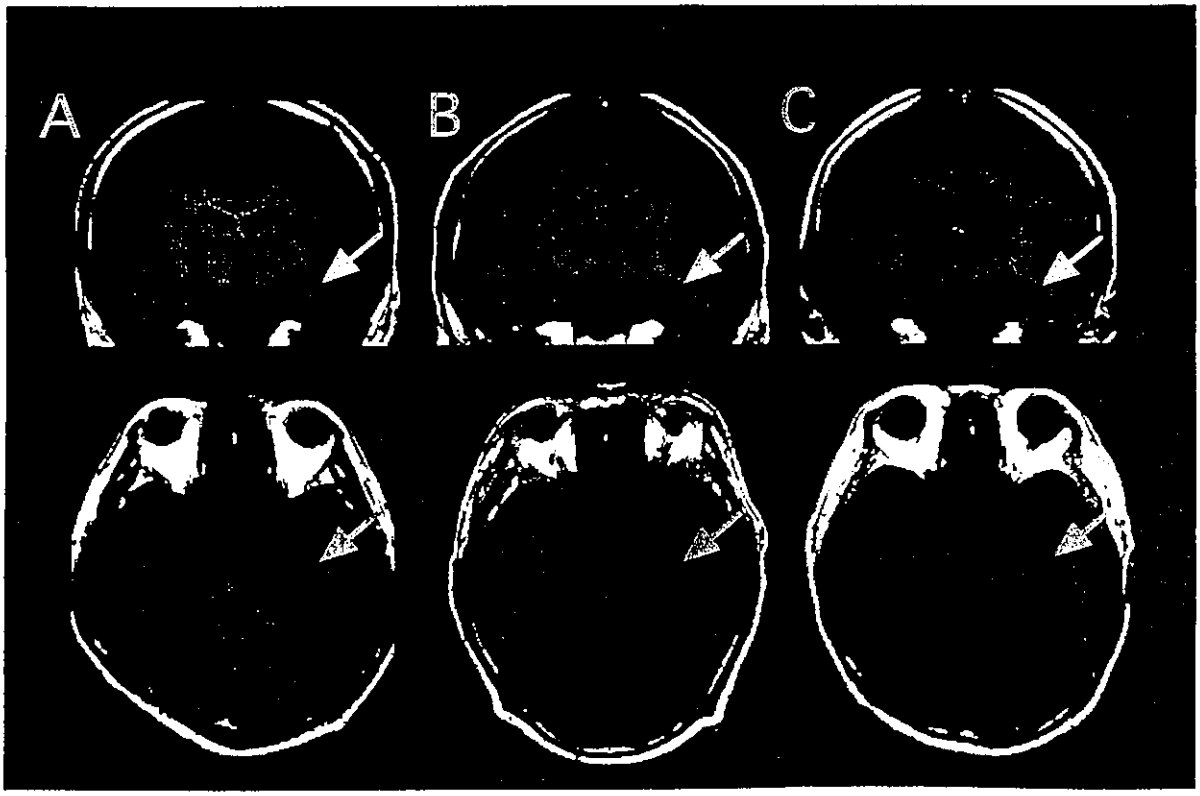
図　3　MRI volumetryにおける内側側頭葉構造の境界

体積測定（volumetry）はMRIの出現後比較的早期から行われ，多くの研究が報告されている。図3にはわれわれがvolumetryに用いている内側側頭葉構造の定義と境界を示す。多くの研究でアルツハイマー型痴呆患者と健常人で海馬の体積に有意な差があるとし，MRIを用いた海馬萎縮の評価に対して診断的価値を認めている。

最近では神経病理学的異常が海馬より早期に認められるといわれている内嗅回（entorhinal cortex）の体積減少を同定することにより，より早期のアルツハイマー型痴呆診断の可能性が研究されている。Juottonenらはアルツハイマー型痴呆群（平均MMSE 20.7）と正常コントロール群の内嗅回と海馬の体積を比較し，内嗅回の体積の場合はアルツハイマー型痴呆の診断の感度が90%，特異度が94%であり，一方海馬ではそれぞれ87%，90%であることを報告している[11]。一方Xuらは健常者，MCI群（平均MMSE 25.7），アルツハイマー型痴呆群（平均MMSE 20.6）の内嗅回と海馬体積を測定し，その診断能力を比較したが，その鑑別能力において内嗅回の有用性は認めなかった[12]。理論上は内嗅回のvolumetryがより優れているはずだが，海馬のvolumetryと差がない結果となったのは，内嗅回の解剖学的同定の難しさが原因と考えられ，今後の課題とするところである。

MRIのvolumetryには特別な技術が必要なことから日常の診療に応用可能な施設は稀であり，通常の臨床場面では海馬萎縮を視覚的に同定しなければならない。図4は健常高齢者とアルツハイマー型痴呆患者のMRI冠状断像を示している。常日頃から痴呆患者を診察していなければ，軽微な海馬体積の減少を視覚的に同定することは困難である。その場合むしろ海馬萎縮によってもたらされる側脳室の下角の拡大のほうが同定しやすいだろう。側脳室下角の拡大の診断的価値については，Frisoniらが MRI冠状断を用いて側脳室下角の幅が軽症アルツハイマー型痴呆と正常被験者とを86%の感度で鑑別することを報告している[13]。また側脳室下角の拡大はCTでもある程度同定可能である。ただし正常圧水頭症では海馬萎縮がなくても側脳室下角が拡大するため注意が必要である。

海馬のvolumetryはアルツハイマー型痴呆と他




A：健常者，B：CDR 0.5のごく軽症アルツハイマー病患者，C：CDR 1の軽症アルツハイマー病患者。上段は冠状断像，下段は水平断像を示す。矢印は側脳室下角を示す。進行とともに海馬は萎縮し側脳室下角は拡大している。

図　4　健常者とアルツハイマー病患者のMRI T1強調画像