EBMでみる鑑別診断に使う検査の臨床研究(1)：対象者の参加方法

寺澤　晃彦
二橋　尚志

はじめに

前回のお話では、悪性膠芽腫（グリオーマ）の陽電子放出断層撮影（PET）検査の例を中心に、検査の研究の基本を駆け足でみました[1]。その第一として、あるテーマの全体を眺める方法であるエビデンスマップを紹介しました。エビデンスマップは、「ある特定分野の研究がこれまでにどのようになっているか？」という課題に対して、既存の研究全体をまとめて報告する研究で、いろいろなテーマ・研究方法が使用されます[2]。そもそも医療の目的は患者さんのアウトカムを改善することが目的です。検査はこれを達成するためにおこなう補助的なものにすぎません。前回ご紹介したエビデンスマップでは「研究目的」と「証拠のレベル」でグリオーマPETの現状をご紹介しました[3]。「証拠のレベル」とは、検査によって患者さんのアウトカムが改善するのをどこまで補助できる証拠があるかを測る「ものさし」のようなものです。具体的には「6つのレベル」があることをご紹介しました[4]。レベルが上がるほど患者さんや社会にとって重要な内容を評価していることになります。

診断精度の研究はこの「ものさし」で「第2のレベル」にあたります（ところで、「診断精度」は状況よって「診断検査精度」や「検査性能」と呼ばれることもあります）。例えば、検査の結果が「陽性」あるいは「陰性」という2つの結果でわかる簡単な例で考えましょう。症状がある患者さんに本当にあるかどうか評価したい体の状態（例えば病気）も1つだけとします。例えば「肺がん検査」で「肺がんがあるかないか」を診断するときなどがこの例に当てはまります。診断精度の研究では、検査が病気の有無を「どれほどうまく見分けられるか」を評価します。慣れない専門用語で少し難しくなりますが、感度・特異度という指標がよく使用されています[5]。感度、特異度ともに0％～100％の数字で検査の性能が表現されます。感度が100％、特異度も100％という検査は病気の有無を完璧に見分けられる検査です。感度が50％、特異度も50％という検

査はサイコロを転がして奇数と偶数のどちらの目が出るかを予想する場合と同じです。検査結果によらず、病気を区別できる能力は「五分五分」ということです。検査を受ける前に病気の可能性が「五分五分」と言われた場合、検査を受けた後でも「五分五分」のままでは困りますね。前回のお話でみていた脳腫瘍（グリオーマ）のエビデンスマップでは、これまでに行われた研究の大多数はこの「第２のレベル」にあるということでした。少しおさらいが長くなりました。今回は鑑別診断に使用する検査についての臨床研究のお話です。２部に分けて鑑別診断とは何か、そしてそこで利用する検査精度研究のデザインをみていきます。この第１部では「レビー小体型認知症[6]」という認知症を起こす病気を診断する検査について私たちが行ったシステマティックレビュー[7]を用いて対象者の参加方法が研究結果に及ぼす影響をお話していきます。

鑑別診断

病気を診断する際に「鑑別診断」という医学の言葉があります。いきなり難しそうな専門用語が出てきました。患者さんは気になる症状を抱えて病院を受診します。例えば今回お話する認知症の症状には「物忘れ」という症状があります。「物忘れ」を起こす病気がたった１つであれば話は単純です。しかし、たくさんの病気が「物忘れ」を起こします。少し難しい病気の名前が出てきます…アルツハイマー型認知症、脳血管性認知症、前頭側頭型認知症、そして今回とりあげる「レビー小体型認知症」。他にもまだたくさん「物忘れ」を起こす病気はあります。このように、症状を起こす可能性がある原因をいくつか考え、その中から可能性がある原因を絞り、最終的な原因をできるだけ特定しようとするプロセスが「鑑別診断」です[8]。それでは「鑑別診断」を最も単純にして見てみましょう。例えば、考えなければいけない原因疾患が５種類ある場合（病気Ａ、病気Ｂ、病気Ｃ、病気Ｄ、病気Ｅ）を考えてみましょう。ここでは５つとも同じ位可能性がありそうなところまでわかっていると仮定します。この状況で本当に便利な検査は「検査Ｘだけ受ければ、病気Ａから病気Ｅの５種類すべてがあるかないか100％わかり、最終診断にたどり着ける」というものでしょう。このような検査があればまさに理想的です。ところが、実際の臨床現場ではこのよう検査は多くはありません。別の検査に「検査Ｙを受ければ病気Ａに限ってあるかないか100％診断できる」というものがあります。例えばある病気に限定的な「診断マーカー」などとよばれる検査がこれにあたります。実際に原因は病気Ａで、この検査Ｙを偶然選んだおかげで病気Ａが診断できたとすればラッキーですが、検査Ｙで病気Ａではないとわかった場合、ここで

話は終わりません。まだ残り 4 つの病気 B から病気 E の可能性が残り、それぞれについては何もわからないわけです。病気 A の頻度が非常に少ない場合（あるいは検査をする前にすでに可能性が低いと予想された場合も同じような状況です）、検査 Y で病気 A だけを評価したところでほとんど鑑別診断は進んでいないこととなります。

鑑別診断を少し別の角度から見てみましょう。診断では病気の頻度だけでなく、病気の重大性や治療法の有無も問題となります。仮に病気 B から病気 E まですべて病気自体はほとんど進行もせず、治療法もないとします。この場合は病気を確定診断しても何も今後の状況は変わらず、診断自体の目的を考えなおすことも必要です。病気 A だけが重大な病気で、しかも治療法がある場合、病気 A だけが評価できればそれで終わりでも良い場合もあります。もちろん、たとえ重要ではない病気 B から病気 E もこのなかのどれが原因かわかることが患者さんの納得や安心というメリットになることもあります。このような場合は診断できるということが違った意味で大事なのかもしれません。

では、病気 A ではないことが分かった状態で、やはり病気 B から病気 D まで残り 4 つすべて調べることが重要な場合はどうでしょうか。単純に考えればそれぞれの病気に対して検査 Y のような病気があるかないかを100％診断できる「診断マーカー」検査が必要そうですね。実際にこのようにすべての病気に対して理想的な検査をしていくのは簡単ではありません。理想的な検査がないかもしれません。考えなければならない病気がもっと多い（たとえば症状を起こす30の病気全部！）場合はどうしましょう。

レビー小体型認知症

さて、「レビー小体型認知症」にお話を戻しましょう。ここからはこの病気のことを英語名 dementia with Lewy bodies の略語である DLB と呼んでいきます。この病気ではレビー小体と呼ばれている病気と関連があるたんぱく質（α シヌクレイン）でできた構造物が頭の神経細胞にたまり、細胞の脱落がおこります[6]。「物忘れ」「幻視（実際にはないものが見えたりすること）」「体の動かしにくさ」などが症状として現れます[9]。病気の原因は完全には解明されていませんし、根治的な治療についてはまだこれからの段階にあります。診断については先ほどの「検査 X」のように、DLB があるかないかを100％診断できる「診断マーカー」があればよいのですが、これについてもまだ開発中というところ

にあります。そこで、現在臨床現場では世界的な臨床診断基準[9]に照らしてこの病気かどうかを診立てるということが行われています。診断が簡単ではないこともしばしばあります。将来的には診断･治療効果判定に役立つ「バイオマーカー」が開発され、さらに病気のメカニズムに効くような画期的な治療が開発されるということが期待されています。このような中で、特に核医学検査を中心とした脳の画像検査と脳脊髄液中のたんぱく質濃度を調べる検査の診断精度研究がたくさん報告されてきています。この報告に基づいて、どれほど診断に役に立つ検査かどうかを研究したのが今回のシステマティックレビューです。

実はこの研究はとても大きなエビデンスレビュー･プロジェクトの一部分で、今回は全部をご紹介できません。というのも、このレビューだけでも DLB 診断に役立つといわれている画像バイオマーカー 4 種類、髄液バイオマーカー 3 種類をまとめているからです[7]。もしご興味があれば研究報告の表 1（エビデンス全体を鳥瞰するテーブル）と図 1（結果のまとめ）をご覧いただければと思います。今回のお話ではイオフルパン[10]を使用したドパミントランスポーターの単一光子放射断層撮影、いわゆる DAT イメージングの結果に絞って認知症鑑別診断目的のバイオマーカー検査研究の難しさを見てみます。DAT イメージングは最新の2005年国際診断基準[9]でも採用されており、日本でも最近保険診療で受けることができるようになりました。

EBM の質問 1 ：研究の内容は信頼できるか？

前回、診断精度の研究がうまくデザインされ、結果が自分の状況に利用できるかを評価するツールに QUADAS-2[11]があるのをご紹介しました。今回もこちらのツールで各研究の実施状況を評価しました。まずどのような状況で研究が行われているか見てみましょう。ここで大事なのは今回のメインテーマである対象者の参加方法です。重要なポイントは 2 つ、①「研究対象者が持っている病気の種類（数）は？」そして②その「対象者の研究へ参加する「門」の数は？」、この 2 点です。

まずは①の対象となる病気の種類（数）です。5 つの研究では DLB を含めて認知症をおこす複数の病気に対して DAT イメージングが行われています。一方、2 つの研究ではすでに国際診断基準でアルツハイマー型認知症（以後 AD と略します）と DLB と診断できた 2 種類の患者さんだけを評価しています。今回の前ふりにもどれば、最初のタイプの研究は「病気 A、病気 B、病気

C、病気 D、…（病気の種類の数だけ続きます）」といった複数の病気が対象となった研究。２番目のタイプは「病気 A と病気 B」２種類だけの研究です。最初の状況はたくさんの病気に対して「鑑別診断」に役立つかもしれません。２番目の状況はひとまず２つの病気に絞られた状況で使うということになります。

では②の患者さんの集め方です。ここでは患者さんが研究に参加するときに「門」をくぐるイメージを想定します[12]。そして「門」の数がいくつあるかを慎重に検討します。まずは、原因となる病気は何かわからないけれど、同じような「認知症（の症状）がある」患者さんを「１つの門」から研究に入っていただき、DAT イメージングを評価することを考えます。実はこのデザインは実際の診療状況と近い状態で、診療所か病院かにより重症度やそれぞれの病気の頻度は変わるかもしれませんが、「病気 A、病気 B、病気 C、病気 D、病気 E…」と複数の病気の患者さんで、似たような症状があり、日常診療でも診断のために検査が必要である１群の患者さんが「１つの門」をくぐって研究に参加します。そこで検査結果が評価できるわけです。これは理想的な研究デザインです。では①の後半の「AD と DLB に限定した研究」ではどうでしょうか。典型的な研究ではまず国際診断基準で AD と診断された患者さんを患者登録などのリストから集めてきます。さらにその中で DAT イメージングを受けていた人に絞って第１の「門」から研究に参加してもらいます。次に国際診断基準で DLB と診断された患者さんを別の患者リストから集めてきます。同様にその中から DAT イメージングを受けていた人に絞って第２の「門」から研究に入っていただきます。診断も検査もすでに終了しており、「後から」結果だけをみることが一般的です。つまりこの研究には患者さんが参加する「門」は２つあるわけで、別の診断を受けた２つの群の患者さんがまるで「症例群」と「対照群」のように別々に参加しているわけです。診断はすでについているわけですから、「典型例」や「診断が容易な患者さん」が多く含まれている可能性があります。このような別々の門から入る「症例群」と「対照群」にはいろいろな背景が違っている可能性があります。この「２つの門は」例えば病気 A、病気 B、病気 C、病気 D、病気 E について「５つの門」（あるいはもっとたくさん）とすることもあります。あれ、病気もたくさんの種類が参加していて、最初にご紹介した理想的なデザインの研究と似ているような気がします？しかし、「５門研究」に参加している５つの病気は、すでに別々の病気が診断されていることがそれぞれの門をくぐる条件となっています。やはり、「典型的」「診断が容易な

患者さん」が含まれやすくなります。このため、「○○門研究」は理想的なデザインの「同様の症状の患者さんが参加した１門研究」よりも診断精度が過大評価（本当の結果より間違ってよい結果がでてしまうこと）されることが多いようです[12-13]。つまり、実際の診療現場では「○○門研究」から得られた結果は本当にそのまま利用してよいのか注意が必要かもしれません。

今回の DAT イメージングのシステマティックレビューでは、理想的に近い「１門研究」デザインが取られていた研究は２つのみでした[7]。残りはすべて「多門研究」のデザインがとられており、ひとまず何らかの診断が確定している患者さんが参加していました。実際に研究の結果を利用する場合には注意が必要です。デザイン上の問題点はほかにもいくつかありました。そのなかでも、認知症をきたす疾患の確定診断は簡単ではないということがあります。何度も出てきた国際診断基準はあくまで臨床診断基準であり、これまでにも出てきたひとつひとつの認知症疾患を完璧には診断できません。患者さんが亡くなった後に脳の病理解剖をして病気の原因を詳しく検査する方法を勧めている専門家もいます。「完璧ではない病気の確定診断検査」が研究結果に及ぼす影響についてはまた別の機会に譲ります。

EBM の質問２：研究は同じような結果が再現されているか？
EBM の質問３：研究の結果はどこで使えるの、どこまで重要なレベルの結果が分かっているの？

DLB と DLB 以外の認知症を見分ける研究は５件（366人）、DLB と AD を見分ける研究は４件（125人）、決して多い数ではありません。この中で、「１門研究」デザインを採用して DLB とそれ以外の認知症を見分ける研究は２件（111人）しかありませんでした。「完璧ではない病気の確定診断検査」の問題も根深い様です。EBM の視点からみてみると、現場でそのまま使える結果の確立についてはまだしばらく先ということになりそうです。

参考文献

１．二橋尚志、寺澤晃彦．健康文．2015年12月（50号）．放射線科学．グリオーマ PET.
２．Miake-Lye IM, Hempel S, Shanman R, Shekelle PG. What is an evidence map ? A systematic review of published evidence maps and their definitions, methods, and products. Syst Rev. 2016 Feb 10；5：28.

3. Nihashi T, Dahabreh IJ, Terasawa T. PET in the clinical management of glioma：evidence map. AJR Am J Roentgenol. 2013 Jun；200(6)：W654-60.
4. Lijmer JG, Bossuyt PM. Various randomized designs can be used to evaluate medical tests. J Clin Epidemiol. 2009 Apr；62(4)：364-73.
5. Altman DG, Bland JM. Diagnostic tests. 1：Sensitivity and specificity. BMJ. 1994 Jun 11；308(6943)：1552.
6. Walker Z, Possin KL, Boeve BF, Aarsland D. Lewy body dementias. Lancet. 2015 Oct 24；386(10004)：1683-97.
7. Mishima A, Nihashi T, Ando Y, Kawai H, Kato T, Ito K, Terasawa T. Biomarkers Differentiating Dementia with Lewy Bodies from Other Dementias：A Meta-Analysis. J Alzheimers Dis. 2015；50(1)：161-74.
8. Sox HC, Higgins MC, Owens DK. Medical Decision Making. 2nd Ed. Wiley-Blackwell. 2015.
9. McKeith IG, Dickson DW, Lowe J, et al. Consortium on DLB. Diagnosis and management of dementia with Lewy bodies：third report of the DLB Consortium. Neurology. 2005 Dec 27；65(12)：1863-72.
10. 日本核医学会. イオフルパン診療ガイドライン初版 2014.
11. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, Leeflang MM, Sterne JA, Bossuyt PM；QUADAS-2 Group. QUADAS-2：a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011 Oct 18；155(8)：529-36.
12. Rutjes AW, Reitsma JB, Vandenbroucke JP, Glas AS, Bossuyt PM (2005) Case-control and two-gate designs in diagnostic accuracy studies. Clin Chem 51, 1335-1341.
13. Whiting, PF, Rutjes, AW, Westwood, ME, Mallett, S：A systematic review classifies sources of bias and variation in diagnostic test accuracy studies. J Clin Epidemiol. 2013；66：1093-1104.

（藤田保健衛生大学救急総合内科准教授）
（名古屋大学医学部放射線科講師）