放射線科学

不完全な参照基準と診断精度

寺澤　晃彦

患者さんの正しい診断に放射線画像検査は欠かせない検査の一つです。検査の性能を表す指標として診断精度が使用されていることを以前のシリーズ[1]でお話ししました。感度・特異度、あるいは尤度比や適中率などの性能の程度を表す「ものさし」を用いたお話しでした。今回お話しする内容はその続きとなります。

診断精度の研究では、診断対象となる病気がありそうな患者さんに全員参加していただき、診断精度を評価したい検査を全員に受けて頂きます。その後、この検査とは全く別の検査や方法でその診断対象となる病気が実際にあるかどうかを確認します。この確認検査のことを reference standard と呼び、「参照基準」などと邦訳されています。

例えば次項の表のように、急性の右下腹部痛がある200人の患者さんの集団を考えてみます。ここで、見逃してはいけない診断対象とする病気を急性虫垂炎であるとします。この例では計算を簡単にするために、虫垂炎がある人を100人、虫垂炎ではない人を100人（つまり特定の病気がある頻度を表す有病率としては50％）とした集団を仮定しましょう。ここで評価したい検査は虫垂炎を診断する腹部単純 CT です。CT 検査と何らかの参照基準それぞれで虫垂炎が確認できるかどうかで陽性・陰性と結果を振り分けます。

感度とは実際に対象となる病気がある人の中で検査が陽性になる割合（病気の人をきちんと病気として見つけられるか？）、特異度とは病気がない人の中で検査が陰性になる割合（病気がない人をきちんと病気がないと判断できるか）のことでした。この例の場合、

感度＝90/100＝90％
特異度＝85/100＝85％

と計算できます。
ちなみに検査の不利益を中心に見たい場合には100％から感度・特異度の指標

表

		参照基準		合計
		陽性	陽性	
腹部単純CT	陽性	90人	15人	105人
	陽性	10人	85人	95人
合計		100人	100人	200人

を引き算して

偽陰性率＝100％－感度＝100％－90％＝10％
偽陽性率＝100％－特異度＝100％－85％＝15％

として評価することもできます。偽陰性は症状がある患者さんが対象の場合には、いわゆる「見逃し」具合を考えるのに役立ちます。一方、疑陽性は特に症状がない「健康な人」が対象の健康診断などの場合、「病気でもないのに病気があるかのように結果が出てしまった…」という程度を見ていることになります。

虫垂炎の話に戻りましょう。参照基準は検査の性能を正しく評価したい場合には、確実に対象疾患を評価可能な「完璧な検証方法」であることが望まれます。虫垂炎の場合、例えば開腹手術などで外科的に虫垂を切除し、本当に炎症があるかどうかを病理検査で調べる方法があります。虫垂炎の有無だけが対象であれば、この方法はほぼ100％評価が達成できる参照基準と考えられます。このような完璧な参照基準はしばしばゴールド・スタンダードと呼ばれます[2、3]。診断精度で表現すれば、ゴールド・スタンダードの感度は限りなく100％に近く、特異度も同様に限りなく100％に近い、ということになります。

ところが、実際の臨床現場ではどうでしょう。このような100％を目指せる検証方法が気軽に利用できる状況はむしろ稀です。最近では急性虫垂炎を抗菌治療で内科的に治療をしながら経過をみる方法も選択肢のひとつと考えられます[4]。疑われれば全員に手術をおすすめするということは現実的ではなくなっているでしょう。このような場合、完璧ではない参照基準、いわゆる臨床的参照基準[2、3]で検査の性能を評価することとなります。虫垂を切り取らずに虫垂炎があるかどうか？・・・真実を判断するのはもはや容易ではなくなりました。状況によってはCT検査よりも検査の性能が劣るかもしれない超音波検査の結

果で判断せざるを得ないかもしれません。内科的な治療をおこなって、入院で様子をみた2週間全体の経過（臨床経過）で判断する場合もあるでしょう。症状は無事に改善したとします。本当に虫垂炎があって抗菌治療が効いて治ったのか、あるいは他の心配ない病態、例えば無治療で良好な経過をたどるウイルス性腸炎などが自然に軽快しただけなのか、もう、これは完全に区別はできません。臨床的参照基準とは不完全で常に不確実なものであることを認識している必要があります。

臨床研究では、このような臨床的参照基準は1種類のものが単独で全員に行われる場合もあります。また、問題がある研究では評価をしたい検査の結果に応じて臨床的参照基準とゴールド・スタンダードが振り分けて使用されることもあります。例えば、CTが陽性なら外科的虫垂切除、CTが陰性なら抗菌治療を行いながら臨床経過で判断をする、というようなことです。たくさんの臨床的参照基準を組み合わせて考えることもあります。実際の臨床ではその場その場で組み合わせがアレンジされることもよくあります。あるいはすでに特定の複数の検査を行うことが標準化されており、「臨床診断基準」というものが考えられている場合もあります。複数の検査をたくさん組み合わせると診断の性能がどんどん良くなり、よりゴールド・スタンダードに近づくように感じるかもしれません。この辺りは単純ではなく、複数検査の功罪もあるとされます[5]。こちらについては別の機会があればまた一緒に検討しましょう。

では、実際に不完全である臨床的参照基準で診断精度が計算されるとどんなことが起こっているのか見てみましょう。高血圧には明らかな病的な原因がない本態性高血圧とよばれるものの他に、何らかの原因によって2次的に高血圧となる状態があります。その中に、腎血管性高血圧（腎動脈狭窄症）があります。これは動脈硬化や血管の病気などがあり、腎臓に血液を送る動脈が細くなり、これが原因で血圧に関係するホルモンのバランスが崩れることが高血圧の原因です。次項の図1は腎動脈狭窄症を診断する超音波検査の診断精度[6]（データを一部改変）を表したグラフ（ROCプロット）です。ここでは、臨床参照基準の診断精度を悪くすると、計算される超音波検査の診断精度が理論的にどのように動くかを見ることができます。

上にある四角はゴールド・スタンダードと考えられている造影剤を使用した血管撮影検査で腎動脈狭窄症を確認して超音波検査の診断精度を確認した場合です。感度が83%、特異度が91%と計算されます。ところが、腎臓の機能がすでに低下している場合、造影剤が安全に使用できないこともあります。この場

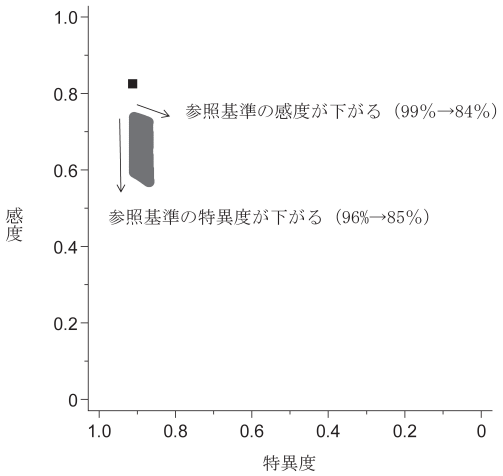

図 1

合、血管撮影検査は行えず、造影剤を使用しない単純 MRI 検査を臨床参照基準として腎動脈を評価するかもしれません。この場合に超音波検査の診断精度はどうなるでしょう。例えば、仮に単純 MRI 検査の診断精度はゴールド・スタンダードを使用して評価すると感度が84％から99％まで、特異度が85％から96％まで、と幅がある値が確認できたとします（実際はもっと精度が低いといわれています）。有病率が25％の集団（つまり、4 人に 1 人に腎動脈狭窄があるハイリスクの集団）で検討したとしてみましょう。数学的なモデル[7]で計算すると、診断精度は先程の四角の下に示した台形のような範囲となります。ゴールド・スタンダードで評価した正しい計算値（上の四角）と比し、診断精度が低く計算されています。精度の悪い検査で評価しているわけですから、取りこぼしをしてしまっているようなイメージです。今回の仮定で最もひどい臨床参照基準で評価した場合には感度が56％、特異度が85％、これは台形の領域で言えば一番右下の頂点・・・上の四角からはずいぶんと離れてしまいました。どうでしょう、不完全な臨床参照基準を用いると、理論上のモデル計算では診断精度が低く見積もられていることが実感できました。

ところが、ここではさらにもう一歩踏み込んでみます。少し厄介なところですが、もう少しだけお付き合いください。実は、上の例で用いた数学的モデルは十全ではありません。超音波検査の結果と単純 MRI の結果がお互いに全く関連していない…と仮定した「独立モデル」から計算しています。しかし、実際には同じような病気のメカニズムから異常の程度を吟味するなど共通した理論的な背景で検査が考えられていることは日常茶飯事です。この状況である検査と臨床参照基準（つまり、もう一つの検査）で対象となる病気の有無を評価

すると、結果の真偽についてはともかく、両者の結果は同じとなる場合が増えます。検査も参照基準も「陽性と陽性」「陰性と陰性」という結果のペアが増えるわけですから、感度も特異度もかなり高く見積もられそうだ…というのが予想できますね。この「検査と参照基準の関連」と「不完全な臨床参照基準」を同時に考えるとなると、全体として診断精度に「どちら向き」の影響が「どの程度」で起こってくるかはもう単純ではありません。

最近の診断精度の臨床研究では、このような複雑なメカニズムによる診断精度への影響をベイズ統計を利用した潜在クラス・モデルで補正する方法がしばしば使われるようになってきました。いくつかの方法がありますが、一次研究[8]だけでなく、メタ解析[9,10]にも応用されています。次項に出てくる図２では、以前のシリーズ[1]でもご紹介した認知症の診断によく用いられる脳血流シンチグラフィーを使用して、レビー小体型認知症とそれ以外の認知症をどれほどうまく鑑別できるかをメタ解析で検討しています。レビー小体型認知症など脳の変性疾患から起こる認知症を確定診断するといえば、ゴールド・スタンダードは脳組織の病理学的検査です。問題は、患者さんが亡くなった後に病理解剖をする以外に簡単には検査ができないというところです。そこで、ほとんどの研究では先ほどご紹介した「臨床診断基準」という複数の検査の組み合わせで診断をおこないます。例えばレビー小体型認知症の「臨床診断基準」はこれまでにも２回改訂が行われ、最近新たな改訂診断基準が提案されたところです[11-13]。診断基準の精度改善が期待されていますが、ここではベイズ潜在クラス・モデルを利用して、精度の値は不確実なものとして補正してみました。３つの丸が集ったグループと周囲の破線で示した領域が３組あります。まずはじめに中央にあるものは臨床診断基準を100％正しいゴールド・スタンダードとみなして計算したものです。もちろん臨床診断基準が100％正しいことはあり得ません。米国では医療用検査デバイスの評価・承認は食品医薬品局で行っていますが、このような不完全な臨床参照基準をもとに評価した検査の指標は「感度・特異度ではなく、臨床参照基準との一致度（陽性・陰性一致度）と（あえて読み直）して提示する」ことを勧めています[14]。通常の医学研究の論文で報告されている感度・特異度の中にはこの一致度に読み替えた方が意味合いとしてわかりやすい状況もたくさんあります。脳血流シンチグラフィーについては私たちのメタ解析でもそのように報告してきました[15]。では、本当の感度・特異度が理論的にはどれほどになりそうなのか、計算値を補正してみましょう。左上のグループは図１で見た検査と臨床参照基準がまったく関連していないと

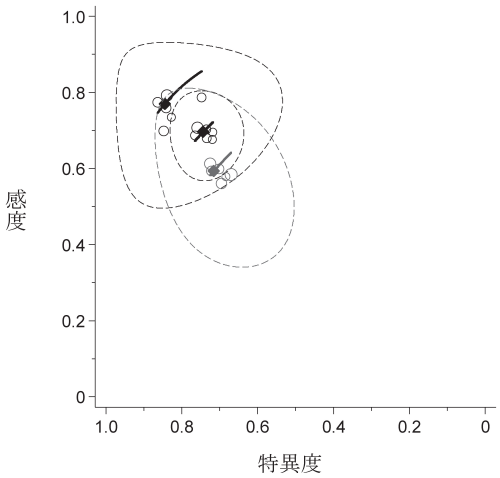

図 2

仮定した簡単な方のモデルです。先ほど実感できた不完全な参照基準による「取りこぼし」、つまり過小評価が補正され、診断精度は大きく改善されています。ずいぶんと左上方にプロットの位置が上がっています。次に、参照基準が不完全ということだけでなく、検査と臨床参照基準に何らかの関連があることも同時に取り込んだモデルが右下のグループです。右下方、つまり診断精度が低い方へと補正されています。こちらのケースではどうやら血流シンチグラフィーと臨床診断基準が同じ結果になる何らかの相関があったようです。繰り返しですが、同じ結果という意味は、結果が「ともに正しい場合」と「ともに間違っている場合」の両者を含みます。

検査と臨床参照基準の関連を取り込まないモデルでは「取りこぼされた」過小評価が改善されますが、関連を取り込むとその状況によっては改善の程度や方向もいろいろになるようです。補正した結果をみながら、検査と参照基準の検査に両者について、検査陽性である「異常な結果」がおこるメカニズムはどうなのか、統計モデルの当てはまり具合はどうか…その他にもいろいろな要素を十分吟味しながら予測される真の診断精度を考える必要がありそうです。

参考文献

1．寺澤晃彦、二橋尚志．健康文．2016年12月（51号）．放射線科学．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.
2．Bossuyt PM, Reitsma JB, Bruns DE et al. STARD 2015：an updated list of essential items for reporting diagnostic accuracy studies. BMJ 2015；351：

h5527.
3. Cohen JF, Korevaar DA, Altman DG, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies : explanation and elaboration. BMJ Open. 2016 ; 6(11) : e012799.
4. Salminen P, Tuominen R, Paajanen H, et al. Five-year follow-up of antibiotic therapy for uncomplicated acute appendicitis in the APPAC randomized clinical trial [e-published September 25, 2018]. JAMA. 2018
5. Dendukuri N, Schiller I, de Groot J, et al. Concerns about composite reference standards in diagnostic research. BMJ. 2018 ; 360 : j5779.
6. Vasbinder GB, Nelemans PJ, Kessels AG, et al. Diagnostic tests for renal artery stenosis in patients suspected of having renovascular hypertension : a meta-analysis. Ann Intern Med. 2001 ; 135(6) : 401-11
7. Trikalinos TA, Balion CM. Options for summarizing medical test performance in the absence of a "gold standard". J Gen Intern Med. 2012 ; 27 Suppl 1 : S67-75.
8. Dendukuri N, Hadgu A, Wang L. Modeling conditional dependence between diagnostic tests : a multiple latent variable model. Stat Med. 2009 ; 28(3) : 441-61.
9. Dendukuri N, Schiller I, Joseph L, Pai M. Bayesian meta-analysis of the accuracy of a test for tuberculous pleuritis in the absence of a gold standard reference. Biometrics. 2012 ; 68 : 1285-1293.
10. Menten J, Boelaert M, Lesaffre E. Bayesian meta-analysis of diagnostic tests allowing for imperfect reference standards. Stat Med. 2013 ; 32 : 5398-5413.
11. McKeith IG, Galasko D, Kosaka K, et al. Consensus guidelines for the clinical and pathologic diagnosis of dementia with Lewy bodies (DLB) : report of the consortium on DLB international workshop. Neurology 1996 ; 47 (5) : 1113-24.
12. McKeith IG, Dickson DW, Lowe J, et al. Diagnosis and management of dementia with Lewy bodies : third report of the DLB Consortium. Neurology 2005 ; 65(12) : 1863-72.
13. McKeith IG, Boeve BF, Dickson DW, et al. Diagnosis and management of dementia with Lewy bodies : Fourth consensus report of the DLB Consortium. Neurology 2017 ; 89(1) : 88-100.

14. Guidance for Industry and FDA Staff. Statistical Guidance on Reporting Results from Studies Evaluating Diagnostic Tests. Document issued on : March 13, 2007.
15. Mishima A, Nihashi T, Ando Y, et al. Biomarkers Differentiating Dementia with Lewy Bodies from Other Dementias : A Meta-Analysis. J Alzheimers Dis. 2015 ; 50(1) : 161-74.

（藤田医科大学救急総合内科学講座）