

EBM でみる鑑別診断に使う検査の臨床研究(2)：結果のいろいろな評価方法

二橋 尚志
寺澤 晃彦

はじめに

本号では鑑別診断に使う検査についてのお話をしています。第1部では鑑別診断の基本をご紹介します¹、レビー小体型認知症を診断する検査について行ったシステマティックレビュー²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ました。第2部では鑑別診断に使う検査が患者さんのアウトカムの改善に役立つ状況が検査精度だけでは評価しにくい状況のお話です。古典的不明熱の鑑別診断を補助する核医学検査のシステマティックレビュー³をご紹介しますながら、検査研究のデザインの難しさをお話していきます。

不明熱

不明熱とは発熱が一定期間続いていて、いろいろ検査をしても原因の確定診断ができないような状況と考えてください。病気の詳しい定義には医学の発展にともない歴史的な変遷があります。最近の権威ある医学書には古典的不明熱の新しい定義が紹介されています⁴。これによると、「もともと健康で抵抗力が落ちるような基礎疾患がない人に起こった、3週間以上継続する38.3℃以上の発熱で、定義に必要な一連の検査を受けたにもかかわらず原因が特定できないもの」という状態です。先ほどの医学書には原因となる病態が100以上も紹介されています。「不明熱の診療は内科医の診断力のみせどころ」などと言われることもありますが、なんとか診断までたどり着いたものでは、感染症、悪性腫瘍、自分に炎症をおこす病気などが診断されています。これ以外の原因もたくさんありますし、診断に至らないことも少なくありません。たとえば別の総説では「よくある」古典的不明熱の原因として38種類の病気を紹介しています⁵。しかし、国や地域、年代、さらには診療や施設の状況など、不明熱の原因となる病気の種類や頻度に影響を与える要素もあります。

このような不明熱の診療は「鑑別診断」の代表でもあります。第1部で鑑別

診断の基本をご紹介しましたが¹、同じような考え方で見てみましょう。古典的不明熱の評価を受けた時点で考えなければいけない原因疾患が20種類ある（病気 A、病気 B、病気 C、病気 D、病気 E…）とします。20種類すべて診断が大事で、治療もある場合には全部の評価が必要となります。前回と同じように、ある病気に対して限定的な「診断マーカー」を利用する状況を想像してみてください。病気 A、病気 B、病気 C、病気 D、病気 E…と20種類すべてに信頼がおける「診断マーカー」があればよいわけです。原因となる病態が100以上も紹介されているとお話ししました。信頼がおける「診断マーカー」が100種類必要な場合もあるのでしょうか。もちろん、このようなたくさんの検査ばかりに頼らないように、ほかの情報も利用して鑑別診断を「絞る」ことを内科診断学は教えています。しかし、不明熱は「必要な一連の検査を受けたにもかかわらず原因が特定できないもの」というわけです。この「必要な一連の検査」は時代の流れとともに種類も増え、性能もどんどん良くなっています。例えば1980年台と2010年台の体幹部 CT 検査を比べると格段に解像度や検査性能が向上しています。この状況で診断がわからない不明熱は本当に困りものです。ここで前回出てきた「本当に便利な『一網打尽』の検査」という考えが再登場します…「検査 X だけ受ければ、病気 A、病気 B、病気 C、病気 D、病気 E…があるかないかよくわかり、最終診断にたどり着ける」というものです。

核医学の検査でクエン酸ガリウム (67Ga)、フルデオキシグルコース (18F-FDG) という悪性腫瘍や炎症が起きている組織に集まりやすい検査用薬剤があります。このような検査用薬剤が不明熱の原因が存在する組織に集積したところを PET や SPECT 検査で撮影し、熱源のありかを特定します。ここに、組織検査などの確定診断ができる検査で「狙い撃ち」をすればよいわけです。67Ga-シンチグラフィー(あるいは67Ga-SPECT/CT)や18F-FDG-PET/CTを『一網打尽』のためのスクリーニングとして利用した「2段階アプローチ」が俄然注目を浴びてきたというわけです。そこでこれまでの研究報告に基づいて、不明熱の核医学検査がどれほど診断に役に立つ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。

実はこの研究も比較的大きなエビデンスレビュー・プロジェクトの一部としておこないました。合計2000人以上を含む42件の研究で4種類の核医学画像検査を扱ったため、全部の結果を詳しくご紹介することができません。もしご興味があれば研究報告の図2と図4（結果のまとめ）、追補図7（各疾患群での診

断寄与度)、追補図8(核医学検査で評価困難な原因一覧)が斜め読みをする際におすすめの図となります。まずはこちらをご覧くださいと思います。今回のお話では18F-FDG-PETあるいはPET/CTを評価した研究のデザインをみながら、診断精度に加えて検査の貢献度をみる方法をご紹介します。FDG-PETを評価した研究は12件(522人)、FDG-PET/CTを評価した研究は22件(1137人)ありました。

EBMの質問1:研究の内容は信頼できるか?

前回同様、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QUADAS-2⁶で評価しました。前回同様、まず対象者の参加方法を見てみましょう。前回同様、重要なポイントは①「研究対象者が持っている病気の種類(数)は?」②その「対象者の研究へ参加する「門」の数は?」、の2点が重要でした。

まずは①の対象となる病気の種類(数)です。さすがは不明熱の研究です。どの研究もたくさんの種類の不明熱となる原因疾患を抱えた患者さんが参加していました。たとえば、不明熱の原因となりうる「肺外性結核」「悪性リンパ腫」「血管炎」の3つの疾患だけについてすでに診断できた患者さんだけを対象として「後ろ向き」に「3つの別々の門(3門研究)」から参加していただいた…というどこかで読んだばかりのようなデザインはありません。②の門も実際に「不明熱という診断のもとで核医学画像検査をうけた」という状況の方が「1つの門」から参加していました(1門研究)。デザインとしては何となくよさそうな感じでしょうか。ところが、不明熱という大きな「くくり」の診断にいたるまでに受けてきた検査に問題がありそうです。例えば、最新の定義にあるような「必要な一連の検査」が十分にやられていないようです。どんな検査がされているのか十分記載されていない報告もたくさんあります。一般に日常診療で行ったカルテの記載内容を後ろ向きに振り返って調査したものはいろいろな問題があるといわれています。ある人は「×××が疑わしいから○○○の検査はした(だけど、「必要な一連の検査」のその他はあまりやっていない)」、別の△△△が疑わしい人は逆に「□□□の検査はしたけど、○○○はしていない」、あるいは「(ほかの検査はほどほどで)いきなりPET/CTで検査をしまえ」…。状況に応じて主治医の裁量で多様な検査オーダーが出されます。しっかり「必要な一連の検査」をあらかじめ定義しており、18F-FDG-PETあるいはPET/CTまえにきちんと全員におこなっていたタイプの前向き研究はたった

3件しかありませんでした。いくら「1門研究」のデザインが取られていても、「1門」をくぐる患者さんが同じ道から入っていない可能性が残り、研究の結果を利用する場合にはやはり注意が必要です。

ほかにもデザイン上の問題点がありました。不明熱をきたす病気の確定診断が、病気ごとに違っているということがその1つです。ある病気は病理診断や遺伝子検査できちんと診断していたものもありますが、3か月間患者さんの様子をみて最終診断を決めたというものもあります。ひどいものでは、PET/CTの映り方をそのまま診断の根拠にしているものもありました。別の問題では、PET/CTに病気が映った人は「積極的検査」をしたが、映らなかったひとには「しばらく様子を見て」病気かどうかを判断したということもありました。「完ぺきではない病気の確定診断検査」や「検査結果ごとに確定診断検査を変えてしまう」というデザインは少し問題で、診断精度を適切に評価できないバイアスが生じるとされています。これらのバイアスの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？

不明熱の研究は、原因となる病気が結局診断できない人がいる、ということをお話しました。多くの研究は原因の確定診断ができた場合と診断が結局分からなかった場合に分けています。そして、それぞれについて検査結果から感度・特異を計算するという方法を取っていました。しかし、そもそも不明熱として原因となる病気は実際にあるわけです。最終的に診断が分からなかったから病気がなかったことにするというのは、本当の意味で診断精度として評価してよいのか疑問が残ります。検査がどのように役立つかを評価する目的の研究ですが、感度・特異度などの診断精度だけでは実際に検査が役立つ度合いを評価しにくいこともあります。ここで、エビデンスマップで使用した「証拠のレベル」を考え直すことが役立ちます。「証拠のレベル」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」であり、「6つのレベル」があることをご紹介しました⁷。今回は「レベル2」である診断精度から「レベル3」である診断変更への貢献を表す「診断寄与率」が検討されていないかを見てみました。もう少し詳しくお話をすると、「検査をうけた人のうち、どれぐらいの人において検査結果が診断結果

を変えることとなったか」ということを評価します。例えば100人の不明熱の患者さんで考えてみます。不明熱のわけですから、検討する検査（18F-FDG-PET あるいは PET/CT）の実施前は診断が分かる人は0人です。これが、18F-FDG-PET/CT をおこなったところ、70人で熱源のありかが画像に映った人が現れ、さらにこの70人の映った場所に対して生検+病理検査などの詳しいピンポイント検査をしたところ50人で不明熱の原因診断が確定したということが起こります。この場合、診断寄与率は50%となります。もちろん、さらに上のレベルとしては「レベル4」である治療変更への貢献を表す指標である「治療寄与率」もあります。いまお話ししていた原因が確定できたひと50人で、確定した原因疾患の治療法があり、実際に治療が開始できた人が40人。残りの7人は治療法がないウイルス感染症で結局これまで通り対症療法を継続しながら経過をみるだけ、さらに3人は治療法がない悪性腫瘍が原因と分かり治療は開始できず。この場合、治療寄与率は40%となります。

12件報告があった18F-FDG-PET の診断寄与率にはばらつきがありました。結果の再現性はまずまずで、メタアナリシスの結果では要約診断寄与率0.44（95%信頼区間：0.31-0.58）でした。22件報告されていた18F-FDG-PET/CT も結果の再現性・ばらつきに問題は残りますが、要約診断寄与率は0.58（95%信頼区間：0.51-0.64）となりました。検査を実施したことにより、それぞれ約4割、約6割の患者さんで確定診断に結びつくきっかけになったということが分かります。18F-FDG-PET および PET/CT の両者にいえることですが、悪性腫瘍と感染症の患者さんが多い研究で診断寄与率も高いようでした。これは18F-FDG がそのような組織に集まるため当然といえば当然です。残念ながら今回の研究では治療寄与率に関する証拠はほとんどありませんでした。診断ができた人のどれぐらいで本当に治療の変更まで結びついたのかということについては分からないままです。今後は検査前にどのような要因を持っている人に検査がより役に立つのか（ある要因のある人は診断寄与率80%以上など）を検討する研究などが期待されます。

参考文献

1. 寺澤晃彦、二橋尚志. 健康文化2017年12月（51号）. 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. Mishima A, Nihashi T, Ando Y, Kawai H, Kato T, Ito K, Terasawa T. Biomarkers Differentiating Dementia with Lewy Bodies from Other Dementia.

- tias : A Meta-Analysis. *J Alzheimers Dis.* 2015 ; 50(1) : 161-74.
3. Takeuchi M, Dahabreh IJ, Nihashi T, Iwata M, Varghese GM, Terasawa T. Nuclear Imaging for Classic Fever of Unknown Origin : Meta-Analysis. *J Nucl Med.* 2016 Jun 23. pii : jnumed.116.174391. [Epub ahead of print]
 4. Bleeker-Rovers CP, van der Meer JWM. Fever of Unknown Origin. In : Fauci A, Hauser S, Longo D, Jameson J, Loscalzo J, Kasper D, eds. *Harrison's Principles of Internal Medicine.* 19 ed. New York, NY : McGrawHill ; 2015 : 135-142.
 5. Varghese GM, Trowbridge P, Doherty T. Investigating and managing pyrexia of unknown origin in adults. *BMJ.* 2010 ; 341 : 878-881.
 6. 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.
 7. Lijmer JG, Bossuyt PM. Various randomized designs can be used to evaluate medical tests. *J Clin Epidemiol.* 2009 Apr ; 62(4) : 364-73.

(名古屋大学医学部放射線科講師)
(藤田保健衛生大学救急総合内科准教授)