《寄　稿》

日本核医学会「診断用放射性医薬品の臨床評価ガイドライン」の解説 (1)　診断用放射性医薬品の臨床評価の考え方について

佐治　英郎[1]　　大西　良浩[2]
日本核医学会　放射性医薬品臨床評価ガイドライン作成委員会*

(核医学 **46**: 369–375, 2009)

はじめに

近年，新しい診断用放射性医薬品の開発が滞っている．院内サイクロトロン製剤と同一規格品としてのフルデオキシグルコース (^{18}F) を除くと，最近 10 年間に本邦で承認された新医薬品は，2004 年のてんかん焦点診断薬イオマゼニル (^{123}I) だけである．この現象は，医薬品の承認審査の世界的標準化の流れの中で，治療用医薬品だけでなく，診断用医薬品の開発のハードルが高くなったためであり，日本に固有の問題ではない．ただし，日本と欧米とでは，放射性医薬品開発の環境に違いがある．それは，欧米では審査当局が臨床評価の指針[1~4] を示しているのに対して，日本にはそれがないことである．

このことを重視した日本核医学会では，平成 14 年から「放射性医薬品臨床評価ガイドライン作成委員会」を組織し，日本アイソトープ協会医学薬学部会放射性医薬品の臨床評価専門委員会と合同で，「診断用放射性医薬品の臨床評価ガイドライン」の作成を開始した．本邦規制当局とのやり取りを経て最終案を平成 17 年日本核医学会学術総会にて「診断用放射性医薬品の臨床評価ガイドライン案」(以下，本 GL 案という) として公表した[5]．

今般，ガイドライン作成委員会では，臨床評価ガイドライン作成の背景や解決すべき課題を明確にし，本 GL 案に示す診断用放射性医薬品の臨床評価の基本的な考え方に関する理解を深めることを目的として，その解説を核医学会誌に掲載することとした．構成は，以下のとおり予定している．

第 1 回：診断用放射性医薬品の臨床評価の考え方について
第 2 回：欧米の画像診断薬の開発に関するガイダンスについて
第 3 回：日本核医学会　診断用放射性医薬品の臨床評価ガイドラインについて

[1] 京都大学大学院薬学研究科医療薬科学専攻病態機能分析学分野
[2] 特定非営利活動法人 (NPO) 健康医療評価研究機構 iHope International
* 委員長：久保敦司 (慶應義塾大学医学部)
委員：日下部きよ子 (東京女子医科大学)，佐治英郎 (京都大学大学院薬学研究科)，中村佳代子 (慶應義塾大学医学部)，本田憲業 (埼玉医科大学総合医療センター)，遠藤啓吾 (群馬大学医学部)，鈴木豊 (医療法人山中湖クリニック PET センター)，利波紀久 (金沢大学大学院医学研究科)，小泉潔 (東京医科大学八王子医療センター)，宍戸文男 (福島県立医科大学)，橋川一雄 (京都大学大学院医学研究科附属高次脳機能総合研究センター)，間賀田泰寛 (浜松医科大学光量子医学研究センター)，山崎純一 (東邦大学医療センター大森病院)
平成 13 年6 月から平成 17 年 10 月．
受付：21 年 10 月 8 日
別刷請求先：京都市左京区吉田下阿達町 46–29
(〒 606–8501)
京都大学大学院薬学研究科
医療薬科学専攻病態機能分析学分野
佐　治　英　郎

Key words: Diagnostic radiopharmaceuticals, Efficacy, Clinical trials, Guideline for clinical evaluation

第 1 回の本稿では，診断用放射性医薬品の臨床評価の考え方として，画像診断の診断性能評価におけるバイアス (bias；偏り) の排除，および，新しい診断技術の有効性を評価する枠組みについて紹介する．

診断用放射性医薬品の有効性とは

診断用放射性医薬品が有効かどうかを問うとき，そこには 2 種類の意味が含まれている．ひとつは，検査の結果得られる診断が十分に正確であるかどうかであり，もうひとつは，その診断を下すことが患者の健康にとって意味のあるものかどうかである[6]．

前者は，放射性医薬品を用いた検査の結果とゴールドスタンダードとを比較することによって，診断性能として評価される．後者は，患者の立場からみると，診断性能よりも上位の概念となる．具体的には，その検査結果によって治療意思決定が変化するか，さらには，患者予後が変化するかということである．

これら，診断性能評価におけるバイアスの排除と画像診断の有効性評価の枠組みの 2 点は，診断技術の technology assessment の根幹をなすものであり，規制当局による承認審査の根拠となるものでもある．

I. 感度・特異度の測定におけるバイアスの排除

臨床研究の結果は，さまざまな原因によって真の値から偏る (バイアス)．バイアスには大きく分けて選択バイアスと情報バイアスとがある．選択バイアスとは，実際に研究対象となる集団が，本来目的とする集団の正しい代表ではないために生じる偏りをいう．情報バイアスとは，実際に研究対象となる集団からデータを収集するときに起こる偏りをいう．バイアスを避けるためには，研究デザインや測定方法を工夫する必要がある．

診断性能の評価は，検査結果とゴールドスタンダードとを比較して感度と特異度を測定することによって行われるが，ここでも，研究を慎重にデザインし確実に実行しないとバイアスの生じる可能性がある[7~9]．たとえば，ゴールドスタンダードとなる検査が実施されている患者だけを選んで検査の診断性能を推定すると，感度が過大評価される傾向になる．

バイアスへの対処は，診断性能評価に関する研究結果の妥当性 (内的妥当性) を担保するための必要条件であり，本シリーズで今後解説していく欧米のガイダンス[1~4]や日本版の臨床評価ガイドライン[5]にも明記されている．また，最近では研究者の側からも，同じ考え方をもとに，診断性能評価に関する研究論文を発表するときのガイドライン (STARD)[10] が発表されている．

診断性能の評価研究における特に注意すべきバイアスとして，つぎの 3 つが挙げられる．

1. 確認バイアス (verification bias, work-up bias)
2. 読影バイアス
   (test interpretation bias, test review bias)
3. 患者スペクトラム (patient spectrum, case mix)

以下にこれら 3 つのバイアスについて，心筋イメージング (以下，MPI) の診断性能の評価を仮想例として説明する．

1. 確認バイアス

確認バイアスは，選択バイアスの一種であり，有病と無病の確認がとられた患者だけを対象とした研究で生じる．冠動脈狭窄の検出に関する MPI の感度と特異度を，冠動脈造影 (CAG) をゴールドスタンダードとして測定する研究において，臨床現場で冠動脈造影が実施されている患者だけを対象とした場合，どんなバイアスが考えられるだろうか．

この研究では，感度が過大評価され，特異度が過小評価されることになる．確認バイアスのメカニズムは Fig. 1 のごとくである．臨床現場では，MPI が陰性の患者では冠動脈造影が施行されない傾向がある．仮に，MPI が陽性の場合にはすべての患者で CAG が施行されており，一方，MPI が陰性の場合には 1/4 だけの患者で CAG が施行されているとしよう．

母集団は冠動脈狭窄を疑う患者集団であり，本

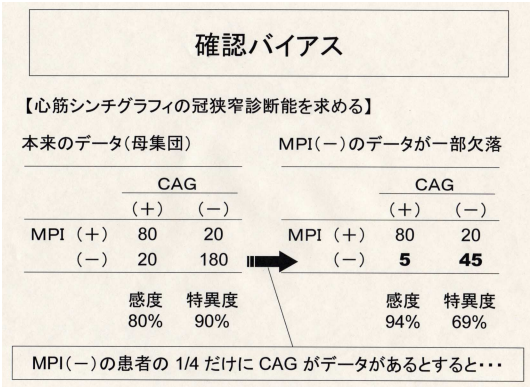

Fig. 1　A numerical illustration of the verification bias.

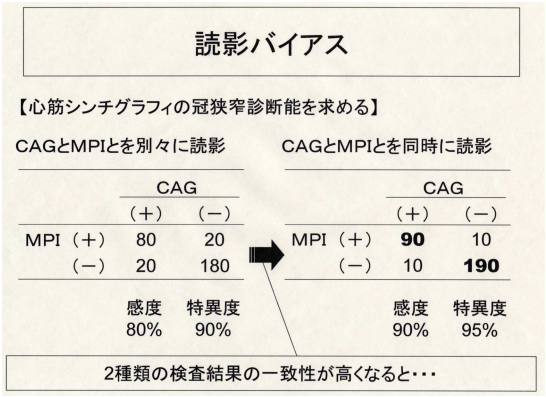

Fig. 2　A numerical illustration of the test-interpretation bias.

来のデータは Fig. 1 左に示される．しかし，この研究では MPI が陰性の患者の 3/4 が除外されており，実際に得られるデータは Fig. 1 右のようになる．本来は 80% であるはずの感度が 94% と過大評価され，本来 90% であるはずの特異度は 69% と過小評価されている．偽陰性および真陰性の患者が選択的に除外されたためである．

逆に，スクリーニング検査の場合に，検査正常で病気の確認がなされていない患者をすべて無病と仮定して，特異度の計算の分母に含めることがあるが，この場合には特異度は過大評価されていることになる．

確認バイアスには，いくつかの特徴がある．第一に，直感に反している．ゴールドスタンダードとの比較が研究の主題であるので，ゴールドスタンダードのない患者を研究対象から除外することは正しいと誤解しがちである．第二に，確認をとるという診療行為が評価対象の検査結果に依存しているほど，バイアスが大きくなる．つまり，よい検査ほど，確認バイアスの影響を受けやすいのである．

確認バイアスを避ける方法は，前向きに研究を行い，研究に登録された全例で確定診断を得ることである．しかしながら，この方法は実施不可能であることが多い．たとえば，がんの病期診断の研究において，すべての部位の病理診断を得ることはできない．確定診断を得ることが困難な場合には，相当の労力を要するものの，経過観察を行うことで確定することが望ましい．そして，すべての患者を対象に統計解析を行う必要がある．

2.　読影バイアス

読影バイアスは，個々の患者において測定を行うときに生じる「情報バイアス」の一種である．画像検査の読影には主観が入る余地があり，それゆえに，そのときに利用できる情報によってバイアスが生じる可能性がある．

再び，MPI の感度と特異度を，CAG をゴールドスタンダードとして，評価する研究を考える．MPI の読影判定を行う評価者が冠動脈造影の結果を知っていたとしたらどうだろうか．ふたつの画像を独立に読影した場合に比べて，所見の一致度が高くなる傾向になると，一般には考えられている (Fig. 2)．つまり，読影が独立に行われていない場合には，診断性能が過大評価されている可能性が否定できないのである．実際，もし，狭窄があった部位を陽性 (血流欠損) と判断しがちになり，狭窄がない部位は血流欠損がない (陰性) と判断しがちになるという読影バイアスの可能性を指摘されたとき，それに反論することは難しいのではないだろうか．

もうひとつのポイントとして，検査結果以外の

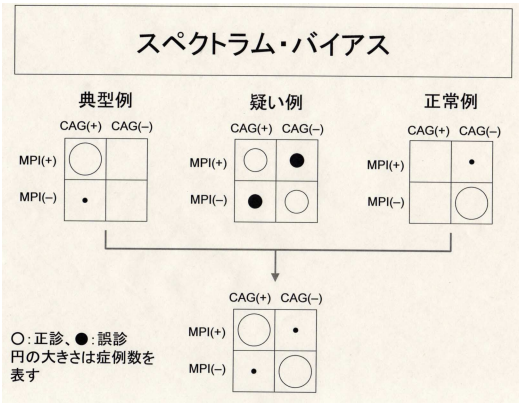

Fig. 3 A schematic illustration of the case-mix.

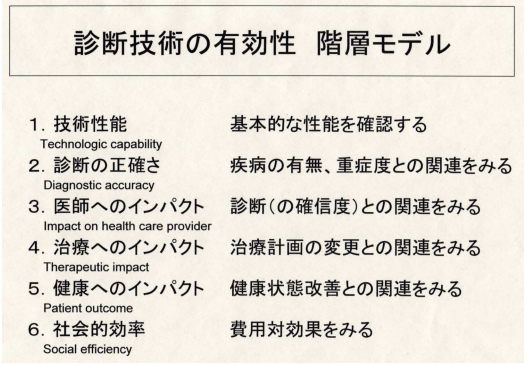

Fig. 4 The hierarchical model of clinical evaluation of diagnostic imaging efficacy.

要素が検査性能に影響することに注意が必要である．中でも最も重要な要素は，患者背景などの臨床情報である．

読影バイアスを避けるためには，読影に際してブラインド化が必要になる．読影者は，少なくとも，ゴールドスタンダードの結果を知っていてはならない．患者に関するその他の情報のうち，何をブラインド化するのかは，個々の研究ごとに決める必要がある．たとえば，臨床情報を含めてすべての情報をブラインド化することができる．また，日常の読影に通常使っているような臨床情報を知った状態で読影することも可能である．画像検査の臨床評価においては，日常現場での状況に近いという理由で，患者背景など通常読影に必要な情報を開示することが一般的であると思われる[1,7]．

3. 患者スペクトラム

ケース混合 (case mix) とも呼ばれる．研究対象となった患者集団の特性によって診断性能の測定値が異なることから生じる偏りのことを指す．

対象疾患の典型的な患者や重症の患者では，ほかの患者と比べて，検査所見が典型的であり，したがって，検査結果もより正確であろう．極端な例として，心筋梗塞の確定した患者と，冠動脈が正常の患者コントロールとを対象として MPI の診断性能を評価するケースを想定しよう (Fig. 3)．おそらく，研究対象には，心筋梗塞疑いで日常遭遇する平均的な患者よりも，MPI 所見の正常・異常がはっきりした被験者が多く含まれているであろう．感度の値は典型的な有病者から得られ，特異度の値は典型的な無病者から得られることになる．このように，診断容易な典型例だけを対象とした研究では，診断性能が過大評価される可能性が高いのである．

もちろん，後に述べるように，診断技術開発の初期段階では，典型例を対象とした研究が効率的である．しかし，臨床現場での診断性能を正しく評価するためには，臨床現場で検査が必要となる患者集団を対象とした研究が必要とされる．

II. 診断技術の有効性評価の枠組み

医療の目的は，最終的には患者の健康の改善である．画像診断はその結果が医師によって解釈され，治療方針の決定に寄与し，結果として患者の健康を改善させる．しかし，このプロセスのすべてをデータとして示すことは，容易ではない．また同時に，このプロセスのどこまでをデータとして示せば画像診断が有効であるといえるのかについてのコンセンサスも十分には成立していない．

1980 年半ば，北米における高額な MR イメージングの医療現場への導入がきっかけとなり，複数の研究者グループによって上記の問題が論じら

れた[11〜16]．これらの研究成果から，「診断技術の有効性階層モデル」が導かれた (Fig. 4)．各階層はより高い階層の必要条件となっている．

本シリーズで今後解説していく欧米のガイダンス[1〜4]や日本版の臨床評価ガイドライン[5]は，「有効性階層モデル」の枠組みの中で，審査の科学性と試験の実施可能性とのバランスを探ったものといえる．この意味で，「有効性階層モデル」は画像診断の臨床評価ガイドラインの鍵となる考え方である．以下に，それぞれの階層について，ここでも主に MPI の開発をモデルにして，順次説明する．

1. 技術的性能

放射性医薬品が，想定した臓器に集積し，患者において診断が可能な画像が得られるであろうことを示す．たとえば，心筋イメージング剤として開発された放射性医薬品の心筋への集積を確認することが挙げられる．

また，その放射性医薬品が診断に役立つ可能性のある疾患や患者の状態を同定することもこの階層に含まれる．たとえば，可能性のある適応疾患として，虚血性心疾患，心筋症，不整脈など様々な疾患の患者に MPI を適用してみることが挙げられる．

この段階で対象とすべき疾患は，病態がよく理解されており，新しい画像検査を用いることで重要な情報が得られると期待されるものである．対象となる患者は，関心のある疾患を有する典型例ということになるであろう．この段階は，その後の試験のために仮説が構築される段階であり，検証的な計画と分析よりも，記述的・探索的な側面が重視される．

2. 診断の正確さ

検査結果がどの程度正確に真の状態を表すかを，感度・特異度などの指標によって測定する．この際の留意点は本稿の I. に紹介したとおりである．

3. 医師に対するインパクト

画像検査の結果は，医師の思考に影響を与えるというプロセスを経て，診療方針の設定に寄与する．医師に対するインパクトは，検査によって診断が変更されることだけではなく，医師の診断に対する確信が高まることをも示す．医師が一連の診察や検査を行ったあとに，追加の画像診断を実施することによって，診断の確信がさらに強固になると考えられる．

しかし，この場合の確信度は主観的・抽象的であり，個々人の経験・考え方・環境等に依存しているかもしれない．実際にはこの診断確信度を定義し，科学的に評価することは困難であるので，診断確信度が臨床試験のエンドポイントとして用いられることは少ない．

4. 治療へのインパクト

新しく導入した検査が，診療方針の決定にどのような影響を与えるかということである．既存の情報をもとに決定した診療方針が，新しい検査の追加によって変更される割合などを指標として評価する．

たとえば，CAG で狭窄を認めたが，典型的な症状がなく経過観察となった患者において，後に MPI によって広範囲の血流低下を認め，心イベントのリスクが高いと判断し血行再建術を行ったとすれば，MPI によって診療方針が変更されたことになる．そして，このようなケースが一定割合以上に発生するならば，MPI には治療上のインパクトがあると結論できるであろう．

FDG-PET 検査の実施によるがん患者の診療方針変更について多くの報告があることは説明には及ばないだろう．このような治療上のインパクトをエンドポイントにした研究は，画像診断の臨床的意義を実証するのに適したものであると考えられる．ただし，さまざまなバイアスが混入するおそれがあるので研究デザインには注意を要する[17,18]．

5. 健康へのインパクト

ある検査を受けた場合に，その検査を受けなかった場合と比較して (または，別の検査をうけた場合と比較して)，患者の予後が改善するかどうかということである．

検査結果から治療方針を決定して治療された患者の予後は，個々の患者の特性やその後の治療によって大きく異なるため，検査の効果を評価するための試験は相当の規模にならざるを得ない．また，検査を受けた場合と受けなかった場合とを厳密に比較するためにはランダム化比較試験 (RCT) が必要となる．したがって，この階層の有効性 (健康へのインパクト) を直接証明するための臨床試験は，実施上のハードルが高く，現実的ではないと思われる[6]．

6. 社会的効率 (経済性)

健康へのインパクトのさらにより上位の階層として，社会的効率 (経済性) が示されている[13,15]．医療経済効果の評価方法論は標準化されていないこともあり，経済性を臨床試験のエンドポイントとして設定することは，困難と言わざるをえない．

まとめ

以上，診断用放射性医薬品の臨床評価の考え方を解説した．欧米の規制当局は放射性医薬品を含む画像診断薬の開発に関するガイダンスを示しているが，それらの有効性評価の方法は，本解説に示した考え方を基礎に構成されている．

診断性能の評価研究におけるバイアスの排除は，研究結果の妥当性 (内的妥当性) を担保するための必要条件である．また，画像診断の有効性階層モデルは，臨床試験のエンドポイントと臨床的意義との関係を理論づける上で有用な考え方である．

本邦においても，開発中および今後開発される診断用放射性医薬品について，診断性能と臨床的意義とを科学的に評価できる臨床試験デザインの作成が課題となり，この課題を達成する助けとして診断用放射性医薬品の臨床評価ガイドラインが作成されている．

引用文献

1) Committee for Proprietary Medicinal Products (CPMP), The European Agency for the Evaluation of Medicinal Products. Points to consider on the evaluation of diagnostic agents. November 2001. Accessed at http://www.emea.eu.int/pdfs/human/ewp/111998en.pdf.
2) U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER) and Center for Biologics Evaluation and Research (CBER). Guidance for Industry. Developing Medical Imaging Drug and Biological Products. Part 1: Conducting Safety Assessments. June 2004. Accessed at http://www.fda.gov/CDER/GUIDANCE/5742prt1.pdf.
3) U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER) and Center for Biologics Evaluation and Research (CBER). Guidance for Industry. Developing Medical Imaging Drug and Biological Products. Part 2: Clinical Indications. June 2004. Accessed at http://www.fda.gov/CDER/GUIDANCE/5742prt2.pdf.
4) U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER) and Center for Biologics Evaluation and Research (CBER). Guidance for Industry. Developing Medical Imaging Drug and Biological Products. Part 3: Design, Analysis, and Interpretation of Clinical Studies. June 2004. Accessed at http://www.fda.gov/CDER/GUIDANCE/5742prt3.pdf.
5) 日本核医学会・放射性医薬品臨床評価ガイドライン作成委員会．診断用放射性医薬品の臨床評価ガイドライン．平成 17 年 8 月 11 日．Accessed at http://www.jsnm.org/system/files/k-42-4-01.pdf.
6) Valk PE. Randomized controlled trials are not appropriate for imaging technology evaluation. *J Nucl Med* 2000; 41: 1125–1126.
7) Begg CB, McNeil BJ. Assessment of radiologic tests: control of bias and other design considerations. *Radiology* 1988; 167: 565–569.
8) Begg CB. Biases in the assessment of diagnostic tests. *Stat Med* 1987; 411–423.
9) Begg CB. Methodologic standards for diagnostic test assessment studies. *J Gen Intern Med* 1988; 3: 518–320.
10) Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al; Standards for Reporting of Diagnostic Accuracy Group. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *Radiology* 2003; 226: 24–28; *BMJ* 2003; 326: 41–44; *Ann Intern Med* 2003; 138: 40–44; *Clin Chem Lab Med* 2003; 41: 68–73; *Am J*

Clin Pathol 2003; 119: 18–22; *Clin Biochem* 2003; 36: 2–7; *Acad Radiol* 2003; 10: 664–669; *AJR Am J Roentgenol* 2003; 181: 51–55; *Ann Clin Biochem* 2003; 40: 357–363; *Clin Radiol* 2003; 58: 575–580; *Croat Med J* 2003; 44: 635–638; *Fam Pract* 2004; 21: 4–10.

11) Guyatt GH, Tugwell PX, Feeny DH, Haynes RB, Drummond M. A framework for clinical evaluation of diagnostic technologies. *CMAJ* 1986; 134: 587–594.
12) Freedman LS. Evaluating and comparing imaging techniques: a review and classification of study designs. *Br J Radiol* 1987; 60: 1071–1081.
13) Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. *Med Decis Making* 1991; 11: 88–94.
14) Kent DL, Larson EB. Disease, level of impact, and quality of research methods. Three dimensions of clinical efficacy assessment applied to magnetic resonance imaging. *Invest Radiol* 1992; 27: 245–254.
15) Houn F, Bright RA, Bushar HF, Croft BY, Finder CA, Gohagan JK, et al. Study design in the evaluation of breast cancer imaging technologies. *Acad Radiol* 2000; 7: 684–692.
16) Evidence-Based Radiology Working Group. Evidence-based radiology: a new approach to the practice of radiology. *Radiology* 2001; 220: 566–575.
17) Guyatt GH, Tugwell PX, Feeny DH, Drummond MF, Haynes RB. The role of before-after studies of therapeutic impact in the evaluation of diagnostic technologies. *J Chronic Dis* 1986; 39: 295–304.
18) Kalff V, Hicks RJ, MacManus MP, Binns DS, McKenzie AF, Ware RE, et al. Clinical impact of (18)F fluorodeoxyglucose positron emission tomography in patients with non-small-cell lung cancer: a prospective study. *J Clin Oncol* 2001; 19: 111–118.