JAMAevidence

Users' Guides to the Medical Literature

A Manual for Evidence-Based Clinical Practice, Second Edition

Edited by Gordon Guyatt, Drummond Rennie, Maureen O. Meade, Deborah J. Cook

医学文献ユーザーズガイド

根拠に基づく診療のマニュアル

第2版

監訳

相原 守夫　池田 正行
三原 華子　村山 隆之

凸版メディア株式会社

USERS' GUIDES TO THE MEDICAL LITERATURE

A MANUAL FOR EVIDENCE-BASED CLINICAL PRACTICE

SECOND EDITION

The Evidence-Based Medicine Working Group.

Edited by Gordon Guyatt, Drummond Rennie, Maureen O. Meade, Deborah J. Cook

医学文献 ユーザーズガイド

根拠に基づく診療のマニュアル

第2版

監訳

相原守夫　相原内科医院　院長

池田正行　長崎大学医歯薬学総合研究科　教授

三原華子　独立行政法人国立がん研究センターがん対策情報センター

村山隆之　御殿場石川病院　医療安全管理室長

Part D

診断
Diagnosis

第 14 章

診断の過程

THE PROCESS OF DIAGNOSIS

W. Scott Richardson, Mark C. Wilson

本章の内容

臨床シナリオ

次のような診断状況を考えてみよう.

1. 43 歳の女性が，左胸部の T3 皮膚分節に集族性有痛性水疱の塊を伴って来院し，あなたは帯状疱疹ウイルスの再活性化による帯状疱疹だと認識した.
2. 78 歳の男性が，高血圧の経過観察で再受診する．彼は，4 ヵ月前の最後の通院以降，10kg 体重が落ちている．患者は食欲減退を訴えたが，それ以外の局在症状はなかった．あなたは，患者の奥さんが 1 年前に亡くなったことを思い出し，可能性としてのうつ病を考えるが，患者の年齢と曝露歴（たとえば喫煙）から，他の可能性も示唆された.

診断への 2 つの補完的アプローチ

臨床研究からの**エビデンス evidence** を使用する臨床診断のための確率論的アプローチは，熟練の臨床医が有力なツールとして活用するパターン認識を補完する（図 14–1 を参照）[1–8]．冒頭のシナリオの最初のケースは，このようなパターン認識がいかにすばやく展開するかを示している.

困難あるいは慣れない状況のためにパターン認識が成立しない場合，臨床医は確率論的な診断思考を採用できる．この場合，臨床医は考えられる一連の診断を列挙し，各診断の確率を推定して検査を行い，検査結果に基づいて各診断の確率を増減させ，最終的には，確実な診断がみつかったという確信に至る [9–14]．第 2 のシナリオは正確な診断のためにこのような確率論的アプローチを必要とするような状況を示している.

確率論的方式を適用するには，解剖学，病態生理学，疾患分類に関する知識を必要とする [11,12,14]．臨床研究からのエビデンスも，最適な診断推論に必要な知識の一つである [15–17]．本章の残りの部分では，臨床研究からのエビデンスがいかに確率論的な診断を円滑にするかについて説明する.

図 14–1

パターン認識 vs 確率論的診断推論

パターン認識	確率論的診断推論
目で見て疾患を認識する	臨床評価から検査前確率を生成する
↓	↓
検査後確率と閾値を比較する （通常，パターン認識は 100%近い確率を示唆し，閾値を超えている）	新情報から検査後確率を生成する （繰り返される場合あり）
	↓
	検査後確率と閾値を比較する

結果の集積が臨床問題を定義する

確率論的方式を用いる場合，臨床医はまず病歴の聴取や身体診察からはじめ，診断の手がかりとなるかもしれない個別の所見を明らかにする．たとえば，第 2 のシナリオでは，食欲不振に関連し，4 ヵ月間で 10kg の体重減少が認められたが，局在症状はなかった．経験ゆたかな臨床医は，しばしば，一連の結果を意味のある集積に分類し，「食欲不振による意図せぬ体重減少」などといったように，症状，身体部位，関係する臓器系に関する簡潔な表現に要約する．このような集合は，**臨床問題 clinical problems** と称されることが多く，鑑別診断への確率論的アプローチの出発点となる[11,18]．

臨床医は可能性のある診断を短いリストに選定する

患者の鑑別診断を考える場合，臨床医はどの疾患を追求するかを決断しなければならない．既知の原因がどれも同等の可能性を持つとし，これらすべてを同時に検査した場合（可能性によるリスト possibilistic list），不要な検査が行われる結果となる．一方，経験ゆたかな臨床医は疾患を厳選し，より可能性が高いと考えられるもの（確率論によるリスト a probabilistic list），診断や治療が行われないまま放置された場合に深刻な事態に至るもの（予後によるリスト a prognostic list），または治療への反応がより良好なもの（実用性によるリスト a pragmatic list）をまず考慮する．個々の患者に対する優先的鑑別診断を賢明に選定するには，これら 3 つの要素（確率論，予後，実用性）すべてを考慮しなければならない．

患者の抱える問題についての，単一の最良の説明を，主仮説または作業診断と呼ぶことがある．第 2 のシナリオでは，臨床医は患者の食欲不振と体重減少の最もありそうな原因を，うつ病と考えた．ほかにも，その可能性尤度や，診断や治療が行われなかった場合の深刻な事態や，治療への反応性から，数個（通常 1 ～ 5 個）のその他の診断を初期評価の中で検討する価値があるかもしれない．原因不明の体重減少については，患者の年齢からして腫瘍の疑いもあり，特に過去の喫煙歴から肺癌の可能性を示唆している．

初期の診断評価では，問題の原因が上記以外にあるとはまず考えられないが，その後，初期仮説が成立しなければ，別の原因も浮上してくるかもしれない．体重減少が認められる 78 歳男性について検討する場合，ほとんどの臨床医は初期鑑別診断として吸収不良を引き起こすような疾患を選択することはないが，検査によってうつ病や癌が除外された場合には，このような仮説に目を向けることがあるかもしれない．

検査前確率を推定することが診断過程を円滑にする

（この患者用の鑑別診断である）研究にふさわしい標的疾患の短いリストをまとめた後，臨床医は

これらの状態の順位付けを行う．診断への確率論的アプローチは，臨床医が短いリスト上の各標的状態の確率である**検査前確率 pretest probability** を推定するのを奨励する（図 14–1）[18,19]．すべての候補診断の確率の合計は 1 に等しくなるべきである．

臨床医はこのような検査前確率をどのようにして推定できるか．一つの方法は，黙示的で，同じ臨床問題を抱えた過去の症例の記憶を利用し，そのような過去の患者に見られた疾患頻度を用いて，その現在の患者の検査前確率の推定を導く．しかししばしば，記憶は不完全で，われわれは特に鮮やかだったり最近の経験によって，そして過去の推論によって，過剰に影響を受けるし，新しいエビデンスを十分重視しない傾向がある．さらに，ある臨床問題についてのわれわれの経験は限られているかもしれない．すべてのこれらの要因のために，臨床医の直感に起因する確率が，**バイアス bias** や**ランダム誤差 random error** にさらされてるままになっている[20–22]．

補完的アプローチでは，研究からのエビデンスを用いて，検査前確率の推定値を導く．これに関連する研究の一つに，同じ臨床問題を持つ患者らが徹底した診断評価を受け，原因として診断された頻度のセットを生み，それを臨床医は最初の検査前確率を推定するために用いることができる（第 15 章「鑑別診断」を参照）．関連する研究の 2 つ目の分類は，臨床決断規則または予測規則を生成する．明確な臨床問題を抱える患者らが診断評価を受け，研究者らは，患者を標的状態の確率が異なる複数のサブグループに分離するような臨床的かつ診断的な検査特性を同定するために統計的手法を用いる（第 17.4 章「臨床予測規則」を参照）．

新しい情報が検査後確率を生み出す

臨床診断は動的プロセスである．新情報の登場によって，標的状態や診断の確率が増減するかもしれない．たとえば，意図せぬ体重減少があった高齢男性の場合，この男性の人生に最近起こった重大な出来事（妻の死亡）から，うつ病が原因である確率が高くなるが，局在性の消化管症状がないことから，消化管疾患の確率は低くなる．尤度比は，新情報によって確率がどの程度変化するかを示す（第 16 章「診断検査」を参照）．

臨床医にとって，経験に基づく直感的推定が検査結果の解釈に役立つこともあるかもしれないが，検査結果によって確率がどの程度増減するかについて確信を持つためには，系統的な研究が必要となる．このような研究にはさまざまな形式のものが考えられるが，最も顕著なものとしては，検査精度に関する個々の **1 次研究 primary studies**（第 16 章「診断検査」を参照）や，これらの検査精度の研究を対象とした**システマティック・レビュー systematic reviews**（第 19 章「エビデンスをまとめる」を参照）があげられる．これらの研究結果の妥当性と適用可能性を吟味すると，各臨床分野に役立つ参考資料として，所見や検査結果の識別力に関する情報を収集できる（第 17.2 章「尤度比の例」を参照）[23,24]．

検査後確率と閾値確率の関係が臨床行動を決める

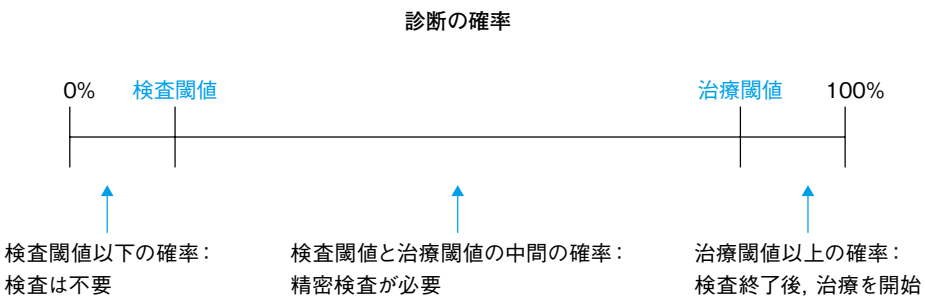
検査結果から検査後確率が生成された後，この新しい確率と 2 つの閾値を比較できる（図 14–2）[25-27]．もし検査後確率が 1 に等しいなら，その診断は確実である．確実性の不足は，検査後確率が 1 に近づくにつれて，診断はますます確実になり，臨床医がその疾患に対する治療を開始することを推奨するであろう確率の閾値（**治療閾値 treatment threshold**）にとどく可能性をますます高める（図 14–2）．これらの閾値は，パターン認識にも，確率論的推論や**ベイズ診断推論 bayesian diagnostic reasoning** にも適用される（図 14–1）．たとえば，ある単一の皮膚分節の領域に有痛性の集族性水疱を呈した 1 つ目のシナリオの患者について考えてみよう．熟練した臨床医は，即座に帯状疱疹の診断をして，患者を治療すべきかを検討するだろう．言い換えると，帯状疱疹の確率はとても高い（ほぼ 1.0, つまり 100%）ので，さらなる検査は必要ない閾値（**治療閾値 treatment threshold**）を超えているからである．

一方，検査後確率が 0 に等しい場合，診断は反証される．0 よりも大きい場合で，検査後確率が 0 に近づくにつれて，それ未満だと臨床医がその診断が除外されるだろうと考える確率閾値（検査閾値）にとどくまで，その診断である可能性はますます低くなる [25]．検査閾値と治療閾値の中間確率では，さらなる検査が求められる．たとえば，それまでは健康だった運動選手が，野球でファウルボールに当たる事故に遭い，胸郭側面に痛みを訴えているとしよう．ここでも，経験を積んだ臨床医なら臨床問題（外傷後の側胸部痛）を認識し，主仮説（肋骨打撲）と他の有効な選択肢（肋骨骨折）を特定し，後者を確認するための検査（レントゲン写真）を計画するだろう．臨床医は，要請があれば，確率が低すぎるためにさらなる検討としない疾患（心筋梗塞など）を列挙することもできるだろう．言い換えると，肋骨打撲の確率ほどは高くはないが，肋骨骨折の確率は検査閾値を上回っているのに対し，心筋梗塞の確率は検査閾値を下回っている．

このような検査閾値や治療閾値は何によって決まるのだろうか．これらの閾値は，検査の特性，疾患の予後，治療の特性に応じて決まる．検査閾値については，検査方式が安全かつ安価で，未診断のまま放置されると深刻な事態に至る疾患で，入手可能な治療が有効かつ安全であるほど，検査

図 14–2

診断過程における検査閾値と治療閾値



閾値が低く設定される．一方，検査方式の安全性に問題があり，コストが高く，未診断でもそれほど深刻でない疾患で，治療の有効性や安全性に確信が持てない場合，検査閾値はより高く設定される．

たとえば，急性冠症候群が疑われる場合にトロポニン検査の指示を行ったとしよう．この疾患は，仮に罹患していれば深刻な帰結（致死性不整脈など）に至る可能性があるが，検査は安価で非侵襲性である．これが理由で，救急科の医師は，急性冠症候群の確率が非常に低い患者にでもこの検査を指示するケースが見受けられる．これはつまり，診断閾値が非常に低く設定されていることを意味する．

これを，肺塞栓症が疑われる場合の肺血管造影と比較してみよう．疾患は深刻だが，検査は侵襲的で，複雑であると考えられる．そのため，ドプラ圧迫超音波検査や換気血流スキャン，ヘリカルCTを経てもなお肺塞栓症の確率が低い場合，臨床医は慎重な経過観察という措置を選択するかもしれない．検査の侵襲性やリスクのために，検査閾値はより高くなっている．

治療閾値については，その後実施される検査が安全かつ安価で，疾患の予後が良好で，治療選択肢のコストや有害作用が大きいほど，閾値は高く設定され，患者に治療を行うにはより確実な診断が必要となる．一方，次に実施する必要のある検査が侵襲的かつ安全性に問題があり，予後が芳しくなく，推奨される治療が安全かつ安価であるほど，確実な診断を下すことよりも治療に着手することの方が望ましいと考えられることから，治療閾値は低く設定される．たとえば，悪性腫瘍の可能性のある患者について考えてみよう．一般的に，臨床医はそのような患者に対して治療を開始する前に，深刻な合併症を伴うかもしれない侵襲的な診断検査を実施しようとするだろう．というのも，治療（手術，放射線療法，化学療法）そのものが罹患率や，場合によっては死亡率に関連するからだ．したがって，臨床医は治療閾値を非常に高く設定する．

これを，胸焼けや胃酸の逆流を訴える患者と比較してみよう．たとえ症状が非定型であっても，臨床医は内視鏡検査を行うのではなく，症状緩和のためのプロトンポンプ阻害薬を処方しようとするだろう．この場合，次の検査の侵襲性と比べた場合の治療の負担が比較的少ないことに鑑みて，治療閾値が低く設定されている．

結論

本章では，診断理論における従来の確率論的教訓について概略し，診断上の決断や措置のために臨床研究からの各種エビデンスを活用する方法について示した．次の章では，診断過程における，ある特定の側面に注目する．

参考文献

1. Elstein AS, Shulman L, Sprafka S. Medical Problem Solving: *An Analysis of Clinical Reasoning*. Cambridge, MA: Harvard University Press; 1978.
2. Schmidt HG, Norman GR, Boshuizen HP. A cognitive perspective on medical expertise: theory and implication. *Acad Med*. 1990; 65 (10): 611–621.
3. Regehr G, Norman GR. Issues in cognitive psychology: implications for professional education. *Acad Med*. 1996; 71 (10 suppl): 988–1001.

4. Redelmeier DA, Ferris LE, Tu JV, Hux JE, Schull MJ. Problems for clinical judgment: introducing cognitive psychology as one more basic science. *CMAJ*. 2001; 164 (3): 358–360.
5. Eva KW. What every teacher needs to know about clinical reasoning. *Med Educ*. 2004; 39 (1): 98–106.
6. Norman G. Research in clinical reasoning: past history and current trends. *Med Educ*. 2005; 39 (4): 418–427.
7. Norman GR, Brooks LR. The non-analytical basis of clinical reasoning. *Adv Health Sci Educ*. 1997; 2 (2): 173–184.
8. Norman GR. The epistemology of clinical reasoning: perspectives from philosophy, psychology, and neuroscience. *Acad Med*. 2000; 75 (10 suppl): S127–S135.
9. Barrows HS, Pickell GC. *Developing Clinical Problem Solving Skills: A Guide to More Effective Diagnosis and Treatment*. New York, NY: WW Norton; 1991.
10. Kassirer JP, Kopelman RI. *Learning Clinical Reasoning*. Baltimore, MD: Williams & Wilkins; 1991.
11. Barondess JA, Carpenter CCJ, eds. *Differential Diagnosis*. Philadelphia, PA: Lea & Febiger; 1994.
12. Bordage G. Elaborated knowledge: a key to successful diagnostic thinking. *Acad Med*. 1994; 69 (11): 883–885.
13. Glass RD. *Diagnosis: A Brief Introduction*. Melbourne, Australia: Oxford University Press; 1996.
14. Cox K. *Doctor and Patient: Exploring Clinical Thinking*. Sydney, Australia: UNSW Press; 1999.
15. Kassirer JP. Diagnostic reasoning. *Ann Intern Med*. 1989; 110 (11): 893–900.
16. Richardson WS. Integrating evidence into clinical diagnosis. In: Montori VM, ed. *Evidence-Based Endocrinology*. Totowa, NJ: Humana Press; 2006: 69–89.
17. Richardson WS. We should overcome the barriers to evidence-based clinical diagnosis. *J Clin Epidemiol*. 2007; 60 (3): 217–227.
18. Richardson WS, Wilson MC, Guyatt GH, Cook DJ, Nishikawa J; Evidence-Based Medicine Working Group. Users' guides to the medical literature, XV: how to use an article about disease probability for differential diagnosis. *JAMA*. 1999; 281 (13): 1214–1219.
19. Sox HC Jr, Blatt MA, Higgins MC, Marton KI, eds. *Medical Decision Making*. Boston, MA: utterworth-Heinemann; 1988.
20. Richardson WS. Where do pretest probabilities come from [editorial, EBM Note]? *Evidence Based Med*. 1999; 4: 68–69.
21. Richardson WS, Glasziou P, Polashenski WA, Wilson MC. A new arrival—evidence about differential diagnosis [editorial]. *ACP J Club*. 2000; 133 (3): A11–A12.
22. Richardson WS. Five uneasy pieces about pre-test probability [editorial]. *J Gen Intern Med*. 2002; 17 (11): 882–883.
23. Fletcher RH, Fletcher SW. *Clinical Epidemiology: The Essentials*. 4th ed. Baltimore, MD: Lippincott Williams & Wilkins; 2005.
24. Straus SE, Richardson WS, Glasziou P, Haynes RB, eds. *Evidence-Based Medicine: How to Practice and Teach EBM*. 3rd ed. Edinburgh, Scotland: Churchill-Livingstone; 2005.
25. Pauker SG, Kassirer JP. The threshold approach to clinical decision making. *N Engl J Med*. 1980; 302 (20): 1109–1117.
26. Gross R. *Making Medical Decisions*: *An Approach to Clinical Decision Making for Practicing Physicians*. Philadelphia, PA: ACP Publications; 1999.
27. Hunink M, Glasziou P, eds. *Decision Making in Health and Medicine: Integrating Evidence and Values*. Cambridge, England: Cambridge University Press; 2001.

第 15 章

鑑別診断

DIFFERENTIAL DIAGNOSIS

W. Scott Richardson, Mark C. Wilson, Thomas G. McGinn

本章の内容

臨床シナリオ

体重減少のある 76 歳男性：どの疾患を探すべきか，またそれらの検査前確率はどれほどか

あなたは，6 ヵ月間で意図せずして 10kg の体重減少があった 76 歳の男性を治療している．今日，長期におよぶ高血圧の追跡のための定期通院で，前回通院時以来体重が減少したと聞かされ，患者は驚きを見せた．患者は，食べる量が減り，食欲がほとんどないとのことだったが，食物に関連した症状はなかった．患者は高血圧のために利尿薬を服用しているが，1 年以上にわたって服用量に変化はなく，時折出現する膝の関節痛や硬直のためにアセトアミノフェンを服用している．喫煙は 11 年前に，飲酒は 40 年前にやめている．検査の結果，やせすぎではあるが，局在性の疾患を示唆する所見はない．初期の血液検査と尿検査の結果は正常である．

意図せぬ体重減少の原因として考えられる数多くの要因を一通り思い浮かべたあなたは，考えられる原因を一度にすべて徹底的に究明することは賢明ではないと考えた．そこであなたは，追究すべき疾患を選択し，それらの疾患の検査前確率を推定するために，意図せぬ体重減少の一般的原因に関する情報をさらに入手したいと考えた．

エビデンスを探す

あなたは，まず，自身の知識不足を疑問として定式化することにした．意図せぬ体重減少が認められる成人が診断評価を受けた場合に，基礎疾患として，どの程度の頻度で腫瘍，胃腸疾患，精神障害などの重要な疾患が認められるだろうか．解決策をみつけるためにコンピュータの前に座ったあなたは，手元に論文からの別冊集を入れたファイルがあるのに気が付いた．あなたは何の気なしに，意図せぬ体重減少に関するファイルを開くと，25 年以上前に出版された，意図せぬ体重減少を経験した患者における一連の疾患の頻度に関する論文が 1 件みつかった[1]．もっと新しいエビデンスをみつけたいと思ったあなたは，まず PubMed にアクセスし，データベースの中でこの古い論文を特定した．「Related Articles」リンクをクリックすると，102 件の引用がみつかった．その中でも 2 番目に新しい，2003 年に出版された Hernandez ら[2] による論文は，体重減少のあった患者における一連の基礎疾患の頻度を明確に取り上げていることから，有用性が高そうであった．さらにリストの下を見ていくと，意図せぬ体重減少に関する最近のナラティブ・レビュー論文がみつかった[3]．そのレビュー論文は，Hernandez ら[2] による論文を，体重減少の原因に関する最新の研究であるとしていた．再確認のために，あなたは電子テキストの中の体重減少に関する章に目を通してみたが，さらに最新の研究については言及されていなかった．最新のエビデンスを特定したことについてある程度の確信を持てたあなたは，批判的吟味のためにその全文を取得した．

ユーザーズガイド

表 15–1 は，鑑別診断のための疾患の確率に関する論文のガイドを要約する．

表 15–1

鑑別診断のための疾患の確率に関する論文のユーザーズガイド

結果は妥当か
　研究患者は，この臨床問題のある患者の全容を代表していたか．
　診断評価は確定的だったか．
結果は何か
　診断および各診断の確率はどれほどか．
　疾患確率の推定値はどれくらい精確か．
結果を患者のケアにどのように適用できるか
　研究患者と臨床セッティングは，自身のものと似ているか．
　このエビデンスが収集されてから，疾患である可能性や確率が変わった可能性は低いか．

結果は妥当か

研究患者は，この臨床問題のある患者の全容を代表していたか

研究対象患者は，検討すべき臨床問題への対応を必要とする基礎疾患を持った標的集団の中から抽出あるいはサンプルとして選ばれる．このサンプルがあらゆる重要な局面において標的集団を反映しているのが理想である．そうであれば，サンプルでみつかった一連の基礎疾患の頻度が，集団全体における疾患の頻度を反映するはずである．標的集団を如実に反映した患者サンプルは，「代表的 representative」と称される．サンプルが代表的であればあるほど，結果として示される疾患確率がより正確になる．表 15–2 に示すように，われわれは，研究対象患者が標的集団全体をどの程度代表しているかを検討する方法を 4 通り提案する．

第 1 に，臨床問題がどう定義されているかを確認しよう．というのも，この定義によって研究対象患者が抽出される標的集団が決まるからだ．たとえば，胸部不快感の研究の場合，研究者の定義に胸部不快感を訴えながらも痛みを否定する患者が含まれていたかどうか（多くの狭心症患者がこれにあてはまる），「胸部」とは前胸部のみの不快感を意味するのか（後胸部はどうか），明らかに最近外傷を負った患者は除外されているかどうかを把握したいと考えるだろう．さらに研究者が，「プライマリケアにおける疲労」といったケアのレベルや[4]，あるいは「原因不明の持続性の咳漱のために紹介を受ける」[5]といった研究に組み入れられる前の評価内容を定義している場合がある．定義が異な

表 15–2

患者サンプルの代表性を確認する

研究者は臨床問題を明確に定義していたか．
研究対象患者は関連する全臨床セッティングから集められたか．
研究対象患者は，臨床セッティングから連続的に組み入れられたか．
研究対象患者は，この問題の臨床症状全般を呈していたか．

れば標的集団も異なり，それによって疾患確率も異なってくる．臨床問題が詳細かつ具体的に定義されていれば，研究のために集めた患者サンプルを，どの標的集団と比較すればよいのかが明確になる．定義が曖昧だと，意図される集団が不明瞭となり，サンプルに含まれる患者が患者全体をどの程度代表しているのか，そしてそのサンプルから得られた疾患の確率の妥当性についての判断に確信が持てなくなる．

第 2 に，患者が組み込まれたセッティングを検討してみよう．同じ臨床問題を抱える患者でも，プライマリケア診療所，救急科，紹介病院など，受診する臨床セッティングはさまざまである．ケアを求めるべきかどうかの選択には，疾患の重症度，各種セッティングの利用可能性，ある臨床医の紹介習慣，患者の好みを含む，いくつかの要因が関わってくると考えられる．これらの影響を考えると，臨床セッティングが異なれば，治療を受ける患者集団の疾患頻度も異なってくると考えられる．一般的に，2 次医療や 3 次医療のセッティングの患者では，プライマリケアセッティングで治療を受ける患者と比べ，重症度が高い，あるいは比較的稀な疾患の割合が高い．たとえば，胸痛を訴える患者の研究では，同じような病歴の患者でも，プライマリケア病院の患者と比べ紹介病院の患者の方が，冠動脈疾患に罹患している割合が高かった[6]．

研究者は，非代表的患者サンプルを治療する可能性が高い特異的セッティングに限定した患者の組み入れは行わないようにすべきである．たとえば，「プライマリケアにおける疲労」の問題についていえば，関連性のあるセッティングはプライマリケアのみだが，研究者は広範なプライマリケアセッティング（例：さまざまな社会経済的地位の患者をケアするセッティング）からの患者組み入れを行うのが理想的である．一般的に，患者の組み入れのために使用された拠点が少ないほど，セッティングが特異的，非代表的である危険性が高い．

第 3 に，研究者が各拠点で患者を特定するために用いた手法，そして患者の見落としを回避するためにどの程度の注意が払われたかに注目してみよう．当該臨床問題の治療のために特定の期間中に研究拠点を受診した全患者の連続サンプルが組み入れられているのが理想である．連続して組み入れられていない場合，基礎疾患が異なる患者が不均衡に組み込まれることになり，サンプルの代表性が損なわれ，疾患確率の妥当性の確信も低くなる．

第 4 に，研究サンプルの患者が示す重症度や臨床的特性の範囲を検討してみよう．軽症，中等症，重症の患者が組み込まれているか．この臨床問題の重要なバリエーションの全てが，このサンプルに含まれているか．たとえば，胸部不快感に関する研究の場合，あなたは，あらゆる重症度の胸痛を呈する患者が組み込まれたかどうか，そして呼吸困難，発汗，放散痛などの重要な関連症状の有無にかかわらず組み込まれたかどうかについて把握したいと考えるだろう．サンプルにおける患者の臨床症状が幅広いものであるほど，より標的集団を代表しているはずである．逆に，臨床症状の幅が狭いほど，サンプルの代表性は低いと評価され，そのサンプルから得られた疾患確率の妥当性についての確信も低くなる．

ユーザーズガイドの適用

Hernandez ら[2] の研究では，臨床問題を，「孤発性の意図せぬ体重減少」と定義していた．これは，局所

性の徴候や症状を伴わず，初期検査では診断に至らなかった，6 ヵ月間で 5%を越える，確認済みの意図せぬ体重減少を意味していた．1991 年 1 月から 1996 年 12 月にかけて，意図せぬ体重減少のために所定の地理的地域から一般内科の外来セッティングおよび入院セッティングへ連続的に紹介された患者は 1,211 人で，そのうち「孤発性」という定義に合致したのは 306 人だった．男女共に組み込まれており，年齢の範囲は 15 ～ 97 歳だった．サンプル患者の人種，文化，社会経済的地位は記載されていない．体重減少が 5kg 未満の患者，意図せぬ体重減少の説明できる診断を過去に受けていた場合，初期評価によって原因が明らかになった場合（例：過去 3 ヵ月以内の利尿薬の使用），体重減少が意図的なものであった場合は，サンプルから除外された．このように，研究サンプルは，意図せぬ体重減少の評価のために紹介を受けた，診断が非常に難しい患者からなる標的集団を非常に適切に代表しており，臨床症状の幅における制約もわずかなものであった．

診断評価は確定的だったか

鑑別診断のための疾患確率についての論文は，研究者らが研究患者に対し，正しい最終診断に到達する時のみ，妥当なエビデンスを提供する．最終診断の正確性 (accuracy) を判断するためには，その診断に至るために使われた診断評価を検討すべきである．この診断評価が確定的であるほど，研究対象者において下された診断の頻度が，標的集団における疾患頻度の正確な推定値を反映している可能性が高くなる．表 15–3 に示すように，診断評価がどの程度確定的かを検討する方法を 6 通り提案する．

第 1 に，研究者の診断評価はどの程度包括的なものか．臨床問題になんらかの原因があるのだとすれば，考えられる原因すべてを検出できる診断評価が理想である．一連の調査が，理にかなった範囲内で包括的であればあるほど，疾患頻度について根拠のない結論が示される確率は低くなる．たとえば，精神状態に変化のあった 127 人の患者における脳卒中の後向き研究では，せん妄を引き起こすあらゆる原因の包括的検索が行われず，118 例は説明のつかないままに終わった[7]．せん妄の原因に関する包括的，系統的検索の説明がないために，疾患の確率の信頼度は低い．

第 2 に，研究患者において実施された診断評価の一貫性を検討してみよう．これは，すべての患者がすべての検査を受けなければならないという意味ではない．むしろ，多くの臨床問題で，臨床医は詳細ながらも的を絞って病歴を聴取・記録し，いくつかの初期検査の実施と同時に，関連する臓器系を対象とした問題重視型の身体診察を行う．その上で，この情報からの診断ヒントにより，いくつかの選択肢のうちの 1 つについて，さらなる検査を実施するかどうかが決まる．すべての患者が同じ初期評価を受け，そこから得られたヒントを基に，所定の検査選択肢が用いられるのが理想的である．確定的な検査結果によって最終診断が確定した時点で，それ以上の検査は不要となる．

所定の診断アプローチを使って患者の疾患が前向きに評価されていれば，その調査が徹底かつ一貫したものであったかを，比較的容易に判断できる．しかし，この判断は，調査が標準化されていない場合には困難である．たとえば，非代償性心不全患者 101 人における増悪因子の研究では，すべての患者に対し，病歴の記録と身体診察が実施されていたが，その後の検査が標準化されていなかったため，疾患確率の正確さを判断するのは困難である[8]．

第 3 に，患者の最終診断を行うために使われる，各疾患に関わる一連の基準を検討してみよう．

診断の可能性のある基礎疾患候補のための一連の明確な基準が作成または採用され，各患者の最終診断に際して一貫してこれらの基準が適用されているのが理想である．可能であれば，これらの基準には各診断を確定するのに必要な所見のみでなく，各診断を除外するのに役立つ所見も含まれるべきである．たとえば，感染性心内膜炎に関する発表されている診断基準には，感染を確認するための基準と感染を否認するための基準とが含まれている[9,10]．このような診断基準があれば，研究者は，複数の病因による症状を呈する患者以外の研究対象患者を，相互排他的な複数の診断群に分類することができる．完全，明確で，参考資料の示された，信用できる一連の診断基準は長文となる可能性があるため，動悸を持つ患者の研究のように印刷論文の別刷りとして[11]，もしくはウェブ版の別刷りとして掲載される場合がある．

診断基準をレビューする際には，「病変の発見 lesion finding」が必ずしも「疾患の説明 illness explaining」と同義ではないことに留意しよう．言い換えると，信用できる診断基準の使用により，研究者らは，臨床問題を説明できるかもしれない複数の疾患に患者が罹患していることを見つけ，そのうち，いずれの疾患が原因であるかについて明らかにできないことがある．疾患の確率の優れた研究は，発見された疾患が実際に患者の病気に寄与していたことを示すなんらかの確証を含むだろう．

たとえば，失神に関する一連の研究において，研究者らは不整脈が原因であると判定する前にはその不整脈と同時に症状が発生していなくてはならないとした[12]．慢性の咳嗽の研究では，研究者らは原因ごとの特定の治療を行い，これに良好な反応を示したときは，この原因が実際に慢性の咳嗽を引き起こしていると確認した[5]．

第 4 に，患者の最終診断が再現可能かどうかを考慮してみよう．再現性を確認するには，まず上述のように，明瞭な基準と包括的かつ一貫した評価を使用する．また，めまいの研究のように[13]，再現性の正式な検査も使用できる．患者への最終診断に対する研究者間の一致が偶然による一致以上であれば，結果として示された疾患確率の妥当性の確信も高まる．

第 5 に，研究の評価にもかかわらず，どれだけの患者が診断に至らなかったかを確認してみよう．包括的な診断評価により，説明のつかない疾患を有する患者が一人もいないのが理想だが，最良の評価を行った場合でもこの目標が達成されない場合がある．未診断の患者の割合が大きいほど，疾患確率の推定値に誤差が生じている確率が高くなる．

表 15–3

確定的な診断評価が行われたか確認する

診断評価は十分に包括的だったか．
診断評価は全患者に一貫して適用されたか．
すべての候補診断のための基準が明瞭かつ信用のできるものだったか．
診断は再現可能か．
診断不明の患者は少なかったか．
診断不明の患者の追跡は十分に長期間で，なおかつ完了していたか．

たとえば，耳鼻科外来の患者 1,194 人における，めまいのさまざまな原因についての後ろ向き研究では，約 27%が未診断のままだった[14]．患者の病気の 1/4 以上は原因不明のため，サンプル全体の疾患頻度は不正確かもしれない．

第 6 に，研究の評価を経てもなお診断がつかない患者がいる場合，患者の追跡期間とその徹底性，そして追加的な診断が行われたか，また臨床アウトカムが把握されているかどうかを確認してみよう．追跡の期間が長く，徹底したものであるほど，研究終了時点で診断がつかなかったにもかかわらず害のなかった患者においては，状態は良性だったという強い確信が持てる．では，どれだけの期間であれば十分なのか．あらゆる臨床問題にあてはまる単一の解答は存在しないが，急性の自然治癒性の症状では 1 ～ 6 ヵ月間，慢性の再発性もしくは進行性の症状では 1 ～ 5 年間が妥当であろう．

ユーザーズガイドの適用

Hernandez ら[2] は，病歴，身体診察，血液検査（血球数，赤沈，血液生化学，蛋白電気泳動，甲状腺ホルモン値），尿検査，レントゲン（胸部と腹部）による標準化された初期評価を一貫して使用し，その後の精密検査は指導医の判断で実施されたことを説明した．各疾患に対する一連の診断基準は列挙されていない．患者の最終診断には，文献内で体重減少の原因として認識された疾患の発見だけでなく，体重減少と疾患の臨床アウトカム（回復または進行）との間の相関も必要だった．診断評価は 2 人の研究者らにより独立して行われ，不一致 (< 5%) は合意により解消された．意図せぬ体重減少を説明できる基礎疾患は，221 人 (71%) の患者で診断され，つまり，85 人 (28%) は最初は未診断であった．追跡期間中，そして 3, 6, 12 ヵ月後の再評価で，85 人中 55 人の患者が診察を受け，41 人の患者が診断された．1 年後の時点で診断不明の患者は 14 人，追跡からの脱落者は 30 人であった．このように，基準が明記されていないことや，追跡からの脱落が 10%あることによる一定の不確実性があるものの，報告されている診断評価は全体としてかなり信用性が高いと考えられる．

結果は何か

診断および各診断の確率はどれほどか

疾患の確率についての多くの研究では，著者は最終診断と各疾患が見つかった患者 y の人数と割合を示した表に主な結果を提示する．その表には，各診断と，各疾患の診断を受けた患者の人数と割合が列挙されている．症状によっては，臨床問題と共存かつ寄与すると考えられる複数の基礎疾患が存在する場合がある．このような場合，著者は患者に対して主診断を同定し，随伴する原因を別の表としてまとめることが多い．あるいは，個別に複数の病因をまとめた集団を同定している研究者もある．

ユーザーズガイドの適用

Hernandez ら[2] は，表中に研究の追跡終了時点で 306 人中 276 人 (90%) の患者に下された診断を示している．たとえば，腫瘍は 104 人 (34%) でみつかり，精神疾患は 63 人 (21%)，原因不明は 14 人 (5%) で同定された．

疾患確率の推定値はどれくらい精確か

妥当なものであったときでも，研究サンプルにおけるこれらの疾患頻度というものは標的集団における疾患の真の確率の推定値にすぎない．これらの推定値の精確さは，著者らによって提示された**信頼区間 confidence intervals (CIs)** によって検討することができる．著者らがそれらを提示していない場合，次の公式を使って自身で信頼区間を計算できる．

$$95\%\ \mathrm{CI} = P \pm 1.96 \times \sqrt{[P\ (1-P)]/\mathrm{N}}$$

ここで，P は，懸案の病因を有する患者の割合で，N はサンプル内の患者数である．この公式は，症例数が 5 人以下の場合は不正確となり，その場合には近似値を使用できる．たとえば，Hernandez ら[2] の研究における，精神科的原因による意図せぬ体重減少の分類について考えてみよう．上記の公式を使うと，$P = 0.23$, $(1-P) = 0.77$, $\mathrm{N} = 276$ から始めることになる．計算すると，CI は 0.23 ± 0.049 であることがわかる．このように，測定された割合は 23%だが，それは 18.1%と 27.9%間で変動するかもしれない．

CI が十分に精確だと考えられるかどうかは，あなたの**検査閾値 test thresholds** または**治療閾値 treatment thresholds** に対して推定された割合と CI がどのような関係にあるかによるだろう．もし推定された割合と 95% CI 全体があなたの閾値の同じ側にあるならば，結果は精確で，検査や治療の計画のために使う疾患確率について確実な結論を可能にする．逆に，もし推定値を取りまく信頼限界があなたの閾値をまたぐならば，結果は十分に精確ではなく，疾患確率について確定的な結論を下すことはできないだろう．妥当だが不精確な確率が示されている場合でも，その不確実性と検査や治療への意味を念頭におきながら，結果を利用することは可能だろう．

ユーザーズガイドの適用

Hernandez ら[2] は，自分たちが見つけ確率の 95% CI は提供していない．前述のように，確率があなたの閾値にどの程度近いかを知りたいならば，自身で 95% CI を計算できる．この場合，CI の下限でさえ意図せぬ体重減少の原因として精神疾患を追究するほど十分に高いようにみえる．

結果を患者のケアにどのように適用できるか

研究患者と臨床セッティングは，自身のものと似ているか

本章の前半で，サンプルの代表性と結果の妥当性を判断するには，研究患者がどのようにして標的集団から選択されたのかを検討するよう強調した．今度は，あなたの患者とあなたの診療への適用可能性という異なる判断をするために，研究サンプルを再度検討すべきである．この疑問を 2 通りの方法で定式化し（「研究患者と臨床セッティングは自身のものと十分に似ており，エビデンスを活用できるほどか」，あるいは，「患者とセッティングは自身のものとはあまりにも異なるため，結果を無視すべきだろうか」），解答を比較してみよう．たとえば，この問題を有するあなたの患者が，特定の疾患が風土病となっているような地域からきているならば，その状態の確率は，風土病のない地域での研究でみられた頻度に比べて格段に高くなることから，あなたの診療への研究結果の適用可能性は制限される．

ユーザーズガイドの適用

意図せぬ体重減少の評価のためにあなたへ紹介された 76 歳男性の場合，Hernandez ら[2] によって述べられた臨床セッティングはかなり良く適合するようにみえる．サンプル患者の部分的説明は，年齢，性別の点でこの男性と十分に似ているようであり，ある程度の不確実性は残るものの，エビデンスが活用できないとするほどの違いはおそらくないだろう．

このエビデンスが収集されてから，疾患である可能性や確率が変わった可能性は低いか

時間の経過とともに，疾患の頻度についてのエビデンスが時代遅れになることがある．古い疾患の中には制圧可能なものがあり，天然痘にいたっては根絶されている[15]．新たな疾患や新たな感染症が生じることもある．このようなことが生じると，考えられる疾患のリストやその可能性が大きく変わり，以前は妥当で適用可能だった研究が，関連性を失ってしまうことがある．たとえば，ヒト免疫不全ウィルスの登場により，全身性リンパ節腫脹，慢性下痢，意図せぬ体重減少などの臨床問題の可能性や確率がいかに劇的に変化したかを考えてみよう．

医学や公衆衛生の進歩の結果として同様の変化が生じることがある．たとえば，原因不明の発熱の研究において，新しい診断技術が，悪性腫瘍を有する患者や不明熱の患者の割合を大きく変えた[16–18]．小児白血病の化学療法のような，生存率を改善する治療の進歩は，治癒後何年も経過した後に 2 次性悪性腫瘍のような合併症を引き起こすかもしれないことから，疾患の可能性に変化をきたすことがある．コレラのような疾患を制圧する公衆衛生対策は，予防された疾患によって引き起こされていたかもしれない臨床問題，この例では急性下痢，におけるその後の病因の可能性を変えてしまう可能性がある．

ユーザーズガイドの適用

Hernandezら[2]の研究は2003年に出版され，研究期間は1991～1997年だった．この場合，あなたの知るかぎりでは，このエビデンスの収集以降，意図せぬ体重減少のあった患者における疾患の原因や確率を変えるような新たな進展はない．

臨床シナリオの解決

意図せぬ体重減少のための評価を受けている76歳男性の例に戻ってみよう．初期評価からはなんの手がかりも得られなかったが，詳細な面接の結果，1年前に妻を亡くして以来，食欲不振を伴う抑うつ気分を示唆する強力な手がかりが得られた．あなたの主仮説は，大うつ病障害が患者の意図せぬ体重減少を引き起こしているというものだが，この診断は，その他の状態を除外するための検査を不要とするほど確実なものではない．Hernandezら[2]の研究から，あなたは有効な選択肢の中に，悪性腫瘍（一般的かつ深刻かつ治療可能）と甲状腺機能亢進症（それほど一般的ではないが，深刻かつ治療可能）を含めることにし，これらの疾患（すなわち，これらの選択肢はあなたの検査閾値を超えている）を除外するための検査を手配した．最後に，研究対象患者のほとんどに吸収不良症候群がなく，あなたの患者にも，意図せぬ体重減少を除いてはこの疾患の特性が認められないことから，これを「その他の仮説 other hypotheses」という分類（すなわち，あなたの検査閾値を下回っている）に入れ，この状態の検査を延期することにした．あなたは，検査前確率の初期推定値として研究からの疾患頻度を使い，その手がかりを基にうつ病の確率を上げ，結果としてその他の疾患の確率は低くなった．

参考文献

1. Marton KI, Sox HC Jr, Krupp JR. Involuntary weight loss: diagnostic and prognostic significance. *Ann Intern Med*. 1981; 95 (5): 568–574.
2. Hernandez JL, Riancho JA, Matorras P, Gonzalez-Macias J. Clinical evaluation for cancer in patients with involuntary weight loss without specific symptoms. *Am J Med*. 2003; 114 (8): 631–637.
3. Alibhai SMH, Greenwood C, Payette H. An approach to the management of unintentional weight loss in elderly people. *CMAJ*. 2005; 172 (6): 773–780.
4. Elnicki DM, Shockcor WT, Brick JE, Beynon D. Evaluating the complaint of fatigue in primary care: diagnoses and outcomes. *Am J Med*. 1992; 93 (3): 303–306.
5. Pratter MR, Bartter T, Akers S, et al. An algorithmic approach to chronic cough. *Ann Intern Med*. 1993; 119 (10): 977–983.
6. Sox HC, Hickam DH, Marton KI, et al. Using the patient's history to estimate the probability of coronary artery disease: a comparison of primary care and referral practices. *Am J Med*. 1990; 89 (1): 7–14.
7. Benbadis SR, Sila CA, Cristea RL. Mental status changes and stroke. *J Gen Intern Med*. 1994; 9 (9): 485–487.
8. Ghali JK, Kadakia S, Cooper R, Ferlinz J. Precipitating factors leading to decompensation of heart failure: traits among urban blacks. *Arch Intern Med*. 1988; 148 (9): 2013–2016.
9. von Reyn CF, Levy BS, Arbeit RD, Friedland G, Crumpacker CS. Infective endocarditis: an analysis based on strict case definitions. *Ann Intern Med*. 1981; 94 (4 pt 1): 505–517.
10. Durack DT, Lukes AS, Bright DK; Duke Endocarditis Service. New criteria for diagnosis of infective endocarditis: utilization of specific echocardiographic findings. *Am J Med*. 1994; 96 (3): 200–209.

11. Weber BE, Kapoor WN. Evaluation and outcomes of patients with palpitations. *Am J Med*. 1996; 100 (2): 138–148.
12. Kapoor WN. Evaluation and outcome of patients with syncope. *Medicine*. 1990; 69 (3): 160–175.
13. Kroenke K, Lucas CA, Rosenberg ML, et al. Causes of persistent dizziness: a prospective study of 100 patients in ambulatory care. *Ann Intern Med*. 1992; 117 (11): 898–904.
14. 14. Katsarkas A. Dizziness in aging—a retrospective study of 1194 cases. *Otolaryngol Head Neck Surg*. 1994; 110 (3): 296–301.
15. Barquet N, Domingo P. Smallpox: the triumph over the most terrible of the ministers of death. *Ann Intern Med*. 1997; 127 (8 pt 1): 635–642.
16. Petersdorf RG, Beeson PB. Fever of unexplained origin: report on 100 cases. *Medicine*. 1961; 40: 1–30.
17. Larson EB, Featherstone HJ, Petersdorf RG. Fever of undetermined origin: diagnosis and follow up of 105 cases, 1970–1980. *Medicine*. 1982; 61 (5): 269–292.
18. Knockaert DC, Vanneste LJ, Vanneste SB, Bobbaers HJ. Fever of unknown origin in the 1980s: an update of the diagnostic spectrum. *Arch Intern Med*. 1992; 152 (1): 51–55.

第 16 章

診断検査

DIAGNOSTIC TESTS

Toshi A. Furukawa, Sharon Strauss, Heiner C. Bucher, Gordon Guyatt

本章の内容

はじめに

本章の前の 2 つの章（第 14 章「診断の過程」，第 15 章「鑑別診断」）において，診断の過程，診断検査の結果によって臨床医の検査閾値や治療閾値の超過がきまることや，正確な**検査前確率 pretest probability** の取得に役立つ研究の使い方を説明した．本章では，臨床医に対して，非常に高い（組み入れ），そして非常に低い（除外）**検査後確率 posttest probabilities** を提供する診断検査について取り上げた論文の活用方法を説明する．本書の後半では，数多くの検査結果を臨床予測規則に統合する論文の活用方法を説明する（第 17.4 章「臨床予測規則」）．

臨床シナリオ

どうすれば認知症を迅速かつ正確に診断できるか

あなたは多忙なプライマリケア医師で，診察患者は高齢者の割合が高い．その日早く，あなたは暮らし向きの良い一人暮らしの 70 歳女性を治療した．診察時，患者は長期間の問題である下肢の関節痛を訴えた．その時あなたは，この患者から心ここにあらずという印象を受けたが，それがどういうことなのかを具体化することができないでいた．記憶や認知機能に関する具体的な質問に対し，患者は以前よりも記憶力が落ちたことは認めたが，それ以外の問題については否定している．時間が迫っていたことから，あなたは変形性関節症の問題に対処し，次の患者の診察に移った．

その晩あなたは，認知障害の可能性が考えられる高齢患者の診察をどうしたら手短に行えるか考えをめぐらせた．ミニメンタルステート (MMSE) についてはよく知っていたが，実施に時間がかかりすぎる．あなたは，より精密な検査を必要とする患者を特定するためにも，認知障害をある程度正確かつ迅速に診断できる簡易ツールがないものか思案した．

エビデンスを探す

あなたはクリニカル・クエスチョンを定式化した．認知障害が疑われる高齢患者において，認知症の診断（またはより精密な検査を必要とする患者の特定）のための簡易スクリーニングツールの精度 (accuracy) はどうか．あなたは，PubMed の Clinical Queries のページから，まず「diagnosis」と「narrow, specific search」を選択した．「dementia AND screen * AND brief」という検索用語で検索を行った結果，48 件の引用がみつかった．過去 5 年間に実施されたヒトを対象とした英語文献に限定すると，21 件に絞られた．認知症の疑いのある患者に焦点をあて，なおかつあなたが過去に用いていた基準である MMSE 並みの精度を報告した論文がないか，抄録に目を通して探した．これら 2 つの基準をいずれも満たしたものとしては，Six-Item Screener (SIS) というツールに関わる結果を報告している論文があった [1]．あなたは論文の全文を電子的に入手し，研究の方法や結果から，自身の診察室で，このツールの使用が正当化されることを望みながら，その論文を読み始めた．

結果は妥当か

表 16–1 は，診断検査の精度について報告する研究の妥当性を評価し，結果を吟味し，適用可能性を判断するためのユーザーズガイドをまとめたものである．

参加患者は診断上のジレンマを呈していたか

診断検査は，それを行わなければ混合してしまうかもしれない状態と障害を鑑別する限りにおいてのみ有用である．ほとんどの検査は，健康な人と重症な人を鑑別することはできるが，この能力は診療では役に立たない．診断が明確である場合には診断検査を必要としないことから，症状が明らかな症例と無症候性の健康な志願者の比較に限定した研究は役に立たない．診療にきわめて類似し，標的状態の軽度な初期徴候を呈した患者が含まれる研究のみが，検査の真価を証明することができる．

表 16–1

診断検査結果の解釈に関する論文のユーザーズガイド

結果は妥当か
- 参加患者は診断上のジレンマを呈していたか．
- 研究者は，その検査を適切かつ独立した参照基準と比較したか．
- 検査と参照基準を解釈する人には，他の結果は盲検化されていたか．
- 研究対象検査の結果にかかわらず，研究者はすべての患者に同じ参照基準検査を実施したか．

結果は何か
- 可能性のある検査結果範囲に関連する尤度比はどれほどか．

結果を患者のケアにどのように適用できるか
- 検査結果の再現性とその解釈は，自身の臨床セッティングにおいて満足のいくものか．
- 研究結果は自身が診察する患者に適用可能か．
- 検査結果は自身の管理戦略を変えるか．
- 検査結果は患者に利益をもたらすか．

結腸直腸癌患者における癌胎児性抗原 (CEA) 検査のストーリーは，誤った患者群の選択によって，診断検査の導入によって高まった期待がいかに打ち砕かれうるかを示している．ある研究では，結腸もしくは直腸の進行癌がわかっている患者 36 人中 35 人で，CEA 上昇がみられた．この値は，正常な人や妊婦，あるいはその他のさまざまな状態の患者でははるかに低かった[2]．この結果は，結腸直腸癌の診断，あるいはそのスクリーニングに対しても，CEA が有用かもしれないことを示唆した．しかし，その後実施された，結腸直腸癌があまり進行していない（したがって，疾患重症度が低い）患者や，その他の癌や胃腸疾患（したがって，大腸癌ではないが混同されやすい疾患）を持つ患者を対象とした研究では，診断ツールとしての CEA 検査の精度は急に下がった．臨床医らは，新たな癌の診断やスクリーニングとしての CEA 測定を適

切に断念した.

診断検査の研究におけるデザインに関連した**バイアス bias** についての 3 つの系統的，経験的調査研究が実施されている．Lijmer ら[3] や Rutjes ら[4] は，診断検査のメタアナリシスを行い，研究デザインのどの局面が検査の見かけ上の診断性能を左右するのかを調べた．Whiting ら[5] は診断検査性能の推定値に与えるバイアスの影響を調べた **1 次研究 primary studies** を系統的に収集し，レビューした.

3 件の研究はいずれも，代表的でない患者の選択に伴うかなりのバイアスを実証していた．標的陽性患者（標的疾患を有する患者，われわれのシナリオでは認知症患者）と標的陰性患者（標的疾患を有さない患者）が別々の集団から組み込まれると，検査の検出力が過大評価される（**相対的診断オッズ比 relative diagnostic odds ratio [RDOR]**: 3.0, 95%信頼区間 [CI]: 2.0 ～ 4.5, RDOR: 1.9, 95% CI: 0.6 ～ 3.73）[3,4]．研究者らが，標的陽性患者と標的陰性患者を同一の集団から組み込むとしても，非連続的な患者サンプリングや後向きのデータ収集は，診断検査性能の推定値を過大評価する可能性がある（それぞれ RDOR: 1.5, 95% CI: 1.0 ～ 2.1, RDOR: 1.6, 95% CI: 1.1 ～ 2.2）[2,3]．われわれは，代表的でない患者選択の研究は範囲バイアスを持つと分類する（第 17.1 章「範囲バイアス」を参照）．表 16–2 は，診断検査の研究におけるバイアスの原因として経験的な裏付けのあるものをまとめたものである.

研究者は，その検査を適切かつ独立した参照基準と比較したか

診断検査の精度（正確さ accuracy）は，「真実 truth」と比較して決定されるのがもっとも良い．研究対象となっている検査を受ける患者全員に適切な**参照 reference**, **基準 criterion**, もしくは**ゴールドスタンダード gold standard**（生検，手術，検死，無治療での長期追跡のような）が適用されていることを，読者は自身で確認しなければならない.

研究がうまくいかないことがあるとすれば，評価対象となっている検査が参照基準の一部となっているときである．検査が参照基準に取り込まれていると，診断検査の検出力を過大評価する可能性が高くなる．したがって，臨床医は満足できる参照基準のための一基準として，独立性を強く主張すべきである.

たとえば，うっ血性心不全の診断のための腹頸静脈逆流の効用を評価した研究を考えてみよう．しかし，この研究は参照検査として，腹頸静脈逆流を含む臨床的，X 線検査基準を使用していた[6]．もう 1 つの例は，末期患者におけるうつ病のスクリーニングツールを評価した研究に由来する．著者らは，うつ病を検出するための単一の質問（「あなたは憂うつですか」）の性能は完全である（感度 ＝ 1.0, 特異度 ＝ 1.0）と主張した．著者らの診断基準には 9 つの質問が含まれ，そのうちの 1 つが「あなたは憂うつですか」だった[7].

診断検査に関する論文を読むとき，もし，あなたが参照基準を受け入れることができないならば

表 16–2

診断精度の研究におけるバイアスの原因に関する経験的エビデンス [a]

	Lijmer ら [3] (RDOR: 95% CI)	Whiting ら [5]	Rutjes ら [4] (RDOR: 95% CI)
参加患者は診断上のジレンマを呈していたか	症例対照デザイン (3.0: 2.0 ～ 4.5)	偏った患者選択（経験的な裏付けあり）	症例対照デザイン (4.9: 0.6 ～ 37.3)
	非連続的な患者選択 (0.9: 0.7 ～ 1.1)		非連続的なサンプリング (1.5: 1.0 ～ 2.1)
	後向きデータ収集 (1.0: 0.7 ～ 1.4)		後向きデータ収集 (1.6: 1.1 ～ 2.2)
研究者らは，その検査を適切かつ独立した参照基準と比較したか		不適切な参照基準（経験的な裏付けあり）	
		混同バイアス（参照基準の一貫としての検査の使用）（経験的な裏付けなし）	混同 (1.4: 0.7 ～ 2.8)
検査と参照基準を解釈する人には，他の結果を盲検化されたか	盲検化なし (1.3: 1.0 ～ 1.9)	レビューバイアス（経験的な裏付けあり）	単盲検あるいは非盲検での解釈 (1.1: 0.8 ～ 1.6)
研究対象検査の結果にかかわらず，研究者はすべての患者に同じ参照基準検査を実施したか	異なる参照検査 (2.2: 1.5 ～ 3.3)	鑑別的検証バイアス（経験的な裏付けあり）	鑑別的検証 (1.6: 0.9 ～ 2.9)
	部分的検証 (1.0: 0.8 ～ 1.3)	部分的検証バイアス（経験的な裏付けあり）	部分的検証 (1.1: 0.7 ～ 1.7)

略語：CI ＝ 信頼区間，RDOR ＝ 相対的診断オッズ比

a　RDOR, 点推定値，95% CI が示される.

（道理にかなった範囲で，つまり，どのみち完全なものはない），その論文は妥当な結果を提供している可能性は低い（表 16–2）[4].

検査と参照基準を解釈する人は，他の結果を盲検化されたか

参照基準が納得のいくものならば，次の問いは，検査や参照基準の解釈者が，その他の調査結果を知っていたかどうか（盲検化された評価）である.

ひとたび臨床医がコンピュータ断層撮影 (CT) スキャンで肺の結節をみると，彼らは胸部 X 線検査でも以前は検出できなかった病変を見ることが可能となる例や，ひとたび心エコーの結果を知ると，彼らは以前には聞こえなかった心雑音が聞こえてくる例を考えてみよう.

参照基準の結果を知ることが検査に影響を及ぼす可能性が高いほど，盲検化された解釈の重要性は高まる．同様に，参照基準が，評価されている検査の知識による解釈の変化に影響を受けやすいほど，参照基準の解釈者を盲検化する重要性は高くなる．Lijmer ら [3] による経験的研究は，程度

は小さいが非盲検化に関連するバイアスを実証しており (RDOR: 1.3, 95% CI: 1.0 ～ 1.9), Rutjes ら[4] は，統計的に有意ではないが矛盾のない RDOR を見つけた (RDOR: 1.1, 95% CI: 0.8 ～ 1.6)（表 16–2）.

研究対象検査の結果にかかわらず，研究者はすべての患者に同じ参照基準検査を実施したか

確認のための参照基準となる検査を患者に受けさせるかどうかが，診断検査の結果によって決まる場合，診断検査の特性が歪められてしまう（検証バイアス[8,9] または精査バイアス[10,11]）.

この状況は 2 通りの場面があると考えられる．第 1 に，評価指標となる検査を受けた特定の患者サンプルのみが参照基準により確認されることがある．たとえば，運動負荷試験の結果が陽性で冠動脈疾患の疑いのある患者は，運動負荷試験の結果が陰性の患者よりも，冠動脈造影（参照基準）を受ける可能性が高いかもしれない．Whiting ら[5] は，この種の，部分的検証バイアスとして知られる検証バイアスのいくつかの記録例をレビューした.

第 2 に，評価指標となる検査の結果が，複数の異なる参照基準で確認されることがある．Lijmer ら[3] や Rutjes ら[4] は，陽性結果の場合と陰性結果の場合とで異なる参照基準を用いることによって大きなバイアスが生じることを見つけた．鑑別的検証バイアスとしても知られるこの種のバイアスの RDOR は，これら 2 つの**システマティック・レビュー systematic reviews** において，それぞれ 2.2(95% CI: 1.5 ～ 3.3)[3]，1.6 (95% CI: 1.9 ～ 2.9)[4] だった（表 16–2）.

肺塞栓症の診断における換気血流スキャンの有用性を評価した Prospective Investigation of Pulmonary Embolism Diagnosis (PIOPED) 研究では，検証バイアスが問題となった．換気血流スキャンの結果が「正常 / ほぼ正常」や「低い可能性」と解釈された患者では，換気血流スキャンの結果がより陽性に近かった患者と比べ，肺血管造影を受ける割合が少なかった（69％対 92％）．臨床医は，肺塞栓症の確率の低い患者に対しては危険を伴う血管造影の実施を躊躇すると考えられることから，これは当然の結果ともいえよう[12].

ほとんどの論文はここで終了し，そのため読者は，換気血流スキャンで「高い可能性」と「低い可能性」とされた患者とでは適切な血管造影を受ける割合が異なっていることから，バイアスの程度は定かでないが，たぶんその程度は大きいだろうと結論しなければならないだろう．しかしながら，PIOPED 研究者は，スキャンの結果が「低い可能性」もしくは「正常 / ほぼ正常」だった 150 人の患者で，血管造影を受けなかった患者（136 人）と血管造影図の解釈が不明瞭であった患者（14 人）に対し，第 2 の参照基準を適用した．つまり，彼らが，治療なしで回復したならば，このような患者は肺塞栓症でないと判断した．したがって，このような患者全員が抗凝固薬による治療を受けずに 1 年間追跡された．追跡期間中に臨床的に明らかな肺塞栓症を発症した患者は一人もいなかった．このことから，換気血流スキャンを受けた時点では，患者にとって重要な肺塞栓症は（もしその後の有害イベント予防のための抗凝固療法を必要とする肺塞栓症と定義するならば）だれにも存在していなかったと結論できる．このように，

PIOPED 研究は，すべての患者に参照基準評価を適用するという目標は達成しているが，全患者に同じ基準を適用していなかった．

ユーザーズガイドの適用

認知障害のための簡易診断検査の研究は 2 つのコホートを含んでいた．一方は，地域在住の 65 歳以上の黒人からなる層化無作為抽出サンプルで，もう一方は Alzheimer Disease Center で認知機能評価を受けるために家族，介護者，医療提供者により紹介された，選別もスクリーニングもされていない連続したサンプルだった．前者のグループには，詳細なスクリーニング検査で認知症の疑いの高い全患者，ならびにその疑いが中程度，あるいは低の患者からなるランダムサンプルが組み込まれた．研究者らは，いずれの集団でも診断の不確実性に直面した．これらの集団は完璧ではない．前者には認知症の疑いの全くない患者が含まれ，後者はすでにプライマリケアレベルにて初期スクリーニングを通過していた（事実，完全版の老年医学的評価を参考にすべきかという疑問は，文献検索のきっかけとなった患者のために解決しようとしている疑問の 1 つである）．幸いにも，2 つの集団において検査特性が似ていたため，懸念材料は大幅に減る．

全患者が SIS を受けていた．SIS では，患者は 3 つの単語（りんご，テーブル，ペニー）を記憶させられ，次に，曜日，月，年を言わされ，最後にヒントなしで 3 つの単語を思い出すよう指示される．結果は，誤回答の数にて示され，0 ～ 6 の値をとる．

参照基準による認知症の診断では，患者は，病歴ならびに身体診察や神経学的所見，MMSE やその他 5 つの検査を含む神経心理学的検査バッテリー，そして参加者の親戚との面談による，老年科精神科医または神経内科医の評価に基づき，「精神障害の診断と統計の手引き（改訂第 3 版）Diagnostic and Statistical Manual of Mental Disorders (Third Edition Revised) [DSM-III-R]」の基準，そして「国際疾病分類，第 10 版 International Classification of Diseases, Tenth Revision (ICD-10)」の基準の双方を満たさなければならなかった．

この参照基準はあなたにとって納得のいくものであったが，SIS や参照診断を行う担当者に，その他の結果がわからないようになっていたかどうかについては，その出版済み論文から把握することができなかった．疑問を解消するために，あなたは筆頭著者に電子メールを送り，詳細を尋ねた．2 通の電子メールのやりとりを経て，神経心理学的バッテリーが「訓練と試験を受けた研究助手」によって実施されたことが明らかになった．一方，参照基準となる診断は，「老年科精神科医，社会心理学者，老人病専門医，神経心理学者からなるコンセンサスチーム」によって行われていた．著者は，「症例については率直な議論が交わされ，チームメンバーは神経心理学的所見を含め，患者の全診療録に自由にアクセスできた」と報告している．SIS に含まれる 6 項目は MMSE に由来するが，「コンセンサスチーム会議で個別のツールとして抜粋されたわけではなかった」．

したがって，盲検化はされていなかったが，それによって重要なバイアスが生じたとは考えられないため，あなたはこの研究の結果について検討してみることにした．

結果は何か

可能性のある検査結果範囲に関連する尤度比はどれほどか

診断検査の結果をどう解釈するかを判断する一環として，患者が標的状態を持つ可能性に関するわれわれの推定が（これを検査前確率と呼ぶ），その診断検査によっていかにより正確な推定となるか（これを標的疾患の検査後確率と呼ぶ）について考えてみたい．ある特定の検査結果の**尤度比 likelihood ratio (LR)** は検査前確率から検査後確率を求めることで可能となる．

シナリオのプライマリケア医師の立場に身を置き，意識明瞭だが認知障害が疑われる 2 人の患者

を考えてみよう．一方の患者は臨床シナリオの中の 70 歳女性で，暮らし向きは良さそうだが，以前より記憶力が落ちたという具体的病状を訴えている．もう一方は 85 歳女性で，やはり長年診察してきた患者であり，今回始めて息子に付き添われて来院した．心配そうな様子の息子が，いつもの朝の散歩で母親が道に迷ってしまったとあなたに告げた．近所の人が自宅から数マイル離れたところで母親をみつけ，その事件について息子に知らせた．母親の元に駆けつけた息子は，部屋の雑然とした様子に驚いた．しかし，診察室での患者は礼儀正しい態度で，その日は調子が悪かっただけで，騒ぎ立てるほどのことはないと訴えた（それを聞いた息子は天井を見上げてあきれた様子を見せた）．あなたが臨床医として直感した認知症の確率，すなわち検査前確率は，これら 2 人の患者で異なっていた．第 1 の女性ではおそらく 20%程度と確率が比較的低く，第 2 の女性ではおそらく 70%程度と比較的高いと考えられた．

正式なスクリーニング検査，すなわちわれわれが例示した SIS は，認知症の有無を確定的に示すというよりはむしろ，その状態の検査前確率を変え，新たな検査後確率を示すものである．この検査前確率から検査後確率への変化の方向と大きさは，検査特性によって決まる．検査特性の中でも最も大きな価値があるのが LR である．

Callahan ら[1] の研究結果を使って LR を説明したい．表 16–3 は Callahan ら[1] の研究における患者コホートの SIS スコアの分布を示したものである．

実際に認知症を持つ患者に 6 という検査結果がでる可能性はどの程度か．表 16–3 によると，認知症患者 345 人中 105 人（あるいは 30.4%）が 6 つとも誤回答であった．さらに，認知症でない患者 306 中 2 人 (0.65%) でも 6 つとも誤回答であったことがわかる．認知症を持つ患者は，認知症を持たない患者と比べてどの程度の確率でこの検査結果（つまり 6 つとも誤回答）を示すだろうか．これを判断するには，つい先ほど計算した 2 つの尤度 (30.4/0.65) の比に注目する必要がある．

表 16–3

認知症とそうでない患者における Six-Item Screener スコア，および各スコアに対応する尤度比

	認知症あり	認知症なし	尤度比
SIS = 6	105	2	47
SIS = 5	64	2	28
SIS = 4	64	8	7.1
SIS = 3	45	16	2.5
SIS = 2	31	35	0.79
SIS = 1	25	80	0.28
SIS = 0	11	163	0.06
合計	345	306	

略語：SIS = Six-Item Screener
データは Callahan ら[1] より転載．

この比は 47 であった．言い換えると，認知症の患者では，そうでない患者と比べ，6 という検査結果は 47 倍起こりやすい．

同様に，検査結果として得られた各スコアに関連する LR を計算することができる．たとえば，5 という検査スコアの LR は (64/345)/(2/306) ＝ 28 である．表 16–3 は各 SIS スコア別の LR を示している．

LR をどう解釈したらよいのか．LR は，得られた診断検査の結果が，どの程度まで標的疾患の検査前確率を上げるか，または下げるかを示す．LR 1 は，検査後確率が検査前確率とちょうど同じであることを意味している．LR が 1.0 よりも大きいと，標的疾患が存在する確率を増加させる．つまり，LR が高いほど，この増加は大きくなる．逆に，LR が 1.0 より小さいと，標的疾患の確率を減少させ，LR が低いほど，確率の減少は大きくなる．

どれくらい大きいと「大きい big」LR, どれくらい小さいと「小さい small」LR なのだろうか．あなたの毎日の診療で LR を使用するうちにあなたなりの解釈のコツはつかめてくるだろうが，次のようなことを大まかな指針にするとよいだろう．

- LR が，＞ 10 あるいは＜ 0.1 の場合，検査前確率から検査後確率へ，大きな，そしてしばしば決定的変化をもたらす．
- LR が，5 ～ 10 あるいは 0.1 ～ 0.2 の場合，検査前確率から検査後確率への，中程度の変化をもたらす．
- LR が，2 ～ 5 と 0.5 ～ 0.2 の場合，確率に小さな（しかし，時として重要な）変化をもたらす．
- LR が，1 ～ 2 と 0.5 ～ 1 の場合，確率にはわずかな（そして，めったに重要でない）変化しか生じない．

LR の大きさと重要性を判断したら，次に，検査前確率から検査後確率を出す際，それらをどのように使うのだろうか．その方法の 1 つとして，検査前確率をオッズに変換し，その結果に LR を乗じて，出てきた検査後オッズを検査後確率に変換することができる．これよりもっと簡単な方法としては，すべての変換作業が自動的におこなわれ，検査前確率から検査後確率を容易に得ることのできる，Fagan[13] により提案されたノモグラムを使うことである（図 16–1）．

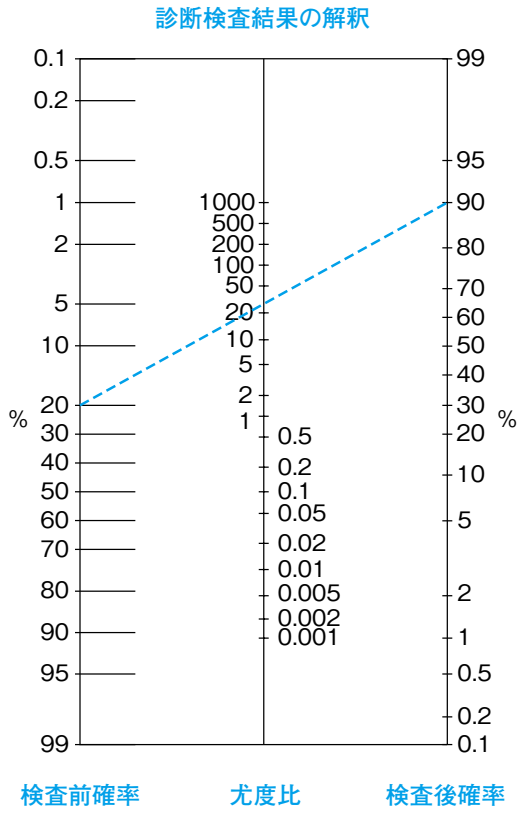
このノモグラムの左目盛りは検査前確率，中央目盛りは LR，右目盛りは検査後確率を表わしている．検査前確率に定規を当て，観測された検査結果に対する LR に回転してあわせると検査後確率を得る．この計算をしてくれるインターネット上の双方向性のプログラム (http://www.JAMAevidence.com) もある．この場合，検査前確率と LR の正確な数値を入力することで，正確な検査後確率が得られる．

冒頭シナリオの，認知症の疑いのある高齢女性を思い出してみよう．われわれはこの患者が認知症である確率は約 20％と判断した．仮にこの患者に SIS で 5 つの誤回答があったと想定しよう．定規をこの患者の検査前確率である 20％に当て，それを 5 という検査結果に関連する LR である 28 にあわせると，検査後確率が約 90％であることがわかる．

検査前確率は，推定値である．鑑別診断に関する文献を参考にして検査前確率を設定できる場合

図 16–1

尤度比のノモグラム



もあるが（第 15 章「鑑別診断」を参照），われわれが知るかぎりでは，認知症の疑いがある場合の検査前確率の直感を補完するような研究はない．直感に頼って検査前確率を精確に推定するのは困難だが，考えられる範囲の検査前確率の意味を検討することで，不確実性に対処することができる．

たとえば，この場合の検査前確率が 10%と低い場合，あるいは 30%と高い場合，ノモグラムを使うと，検査後確率はそれぞれ約 80%，そして 90%以上となる．表 16–4 は，臨床シナリオの 70 歳女性が取得しうる各 SIS スコア別の検査後確率を表形式で示したものである．

この作業を，第 2 の患者，すなわち道に迷った 85 歳女性についてもくりかえすことができる．この患者の病歴や症状からして，認知症の確率は 70%と考えられた．ノモグラムによると（図 16–1），SIS スコアが 6 または 5 の場合の検査後確率はほぼ 100%，SIS スコアが 4 の場合は 94%，SIS スコアが 3 の場合は 85%である．考えられる SIS スコアのそれぞれについて，検査前確率（起こりうる検査前確率の範囲は 60%から 80%），LRs，および検査後確率を表 16–5 に示した．

LRs を使うことを学んだので，あなたは自身の診療で日常的に使用する検査の LRs に簡単にアクセ

表 16-4

認知症の疑いが中程度ある，70 歳女性における検査前確率，Six-Item Screener の尤度比，検査後確率

検査前確率（%）（範囲）[a]	SIS 結果 (LR)	検査後確率（%）（範囲）[a]
20 (10 ～ 30)	SIS = 6 (47)	92 (84 ～ 95)
	SIS = 5 (28)	88 (76 ～ 92)
	SIS = 4 (7.1)	64 (44 ～ 75)
	SIS = 3 (2.5)	38 (22 ～ 52)
	SIS = 2 (0.79)	16 (8 ～ 25)
	SIS = 1 (0.28)	7 (3 ～ 11)
	SIS = 0 (0.06)	1 (1 ～ 3)

略語：LR = 尤度比，SIS = Six-Item Screener

a　カッコ内の値は，考えられる**検査前確率 pretest probabilities** の範囲を表わす．つまり，検査前確率の最良推定値は 20%だが，10 ～ 30%という推定値も考えられる．

スするための資料はないかと考えるかもしれない．JAMA 誌に掲載された Rational Clinical Examination [14] は，病歴や身体診察の診断特性に関する一連の**システマティック・レビュー systematic reviews** をまとめたものである．第 17.2 章「尤度比の例」では，LRs の例をいくつか提示している．さらに多くの例を，ユーザーズガイドのウェブサイト (http://www.JAMAevidence.com) にて蓄積されている．

表 16-5

認知症の疑いが高い 85 歳女性における検査前確率，Six-Item Screener の尤度比 [a]，検査後確率

検査前確率（%）（範囲）[a]	SIS 結果 (LR)	検査後確率（%）（範囲）[a]
70 (60 ～ 80)	SIS = 6 (47)	99 (99 ～ 99)
	SIS = 5 (28)	98 (98 ～ 99)
	SIS = 4 (7.1)	94 (91 ～ 97)
	SIS = 3 (2.5)	85 (79 ～ 76)
	SIS = 2 (0.79)	65 (54 ～ 76)
	SIS = 1 (0.28)	40 (30 ～ 53)
	SIS = 0 (0.06)	12 (8 ～ 19)

略語：LR = 尤度比，SIS = Six-Item Screener

a　カッコ内の値は，考えられる**検査前確率 pretest probabilities** の範囲を表わす．つまり，検査前確率の最良推定値は 70%だが，60 ～ 80%という推定値も考えられる．

連続的検査スコアの 2 値化，感度と特異度，LR＋（陽性尤度比）と LR－（陰性尤度比）

ここまで読み進めてきた読者は，診断検査の解釈の本質的事項を理解したことなる．ただし，**感度 sensitivity** と**特異度 specificity** という 2 つの用語は依然として広く使用されていることから，診断検査に関わる用語として理解しておくことの有用性は高い．診断検査に関する多くの論文は，表 16–6 のような 2×2 表と関連する感度と特異度，そしてこれに対応する図の中で診断検査の全体的性能を提示している（**受信者動作特性 receiver operating characteristic [ROC] 曲線 curve** と呼ばれる）．

Callahan ら[1] の研究では，認知症の診断のためのカットオフ値を 3 つ以上の誤回答とすることを推奨している．表 16–7 は，紹介された患者コホートをこのカットオフ値にしたがって分類したものである．

カットオフ値を 3 以上に設定した場合の SIS の感度は 0.81 (278/345), 特異度は 0.91 (278/306) である．また，表 16–3 と全く同じ方法で LRs を計算することもできる．SIS が 3 以上の場合の LR は (278/345)/(28/306) ＝ 8.8, 3 未満の場合の LR は (67/345)/(278/306) ＝ 0.21 である．陽性検査結果の LR は LR＋と示され，陰性検査結果の LR は LR－と示されることが多い．

では，この 2 値化された 2×2 表を使って，われわれの臨床シナリオの解決を試みよう．
冒頭のシナリオの女性における検査前確率は 20%で，誤回答が 5 つあるものと想定した．SIS スコアの 5 は，LR＋の 8.8 と関連していることから，Fagan のノモグラム[13] を使うと，検査後確率は約 70%となるが，これは，5 つの誤回答に対する固有の LR による検査後確率である 90%よりも大幅に低い数値である．これは，SIS スコアが 3 以上の場合の 2 値化 LR＋は，SIS スコアが 3, 4, 5, 6 の

表 16–6

2×2 表を用いた診断検査の結果と参照基準の結果の比較

		参照基準	
検査結果		**疾患あり**	**疾患なし**
検査陽性		真陽性(TP)	偽陽性(FP)
検査陰性		偽陰性(FN)	真陰性(TN)
感度(Sens)	$= \frac{TP}{TP+FN}$		
特異度(Spec)	$= \frac{TN}{FP+TN}$		
検査陽性の場合の尤度比(LR+)	$= \frac{Sens}{1-Spec} = \frac{真陽性率}{偽陽性率} = \frac{TP/(TP+FN)}{FP/(FP+TN)}$		
検査陰性の場合の尤度比(LR−)	$= \frac{1-Sens}{Spec} = \frac{偽陰性率}{真陰性率} = \frac{FN/(TP+FN)}{TN/(FP+TN)}$		

略語：FN ＝ 偽陰性，FP ＝ 偽陽性，TN ＝ 真陰性，TP ＝ 真陽性
　感度とは，標的状態を持つ患者のうち，検査陽性の患者の割合のことを指す．特異度とは，標的状態を持たない患者のうち，検査陰性の患者の割合のことを指す．

表 16–7

推奨されるカットオフ値を用いた診断検査 (Six–Item Screener) の結果と参照基準（DSM–IV および ICD–10 による協議診断）の結果の比較

	認知症あり	認知症なし
SIS ≧ 3	278	28
SIS ＜ 3	67	278
合計	345	306

略語：DSM–IV ＝ 精神障害の診断と統計の手引き，第 4 版 Diagnostic and Statistical Manual of Mental Disorders (Fourth Edition), ICD–10 ＝ 国際疾病分類，第 10 版 International Classification of Diseases, Tenth Revision, SIS ＝ Six–Item Screener

層を統合するため，その結果として，LR が隣接する層によって希釈されてしまうからである．

この臨床シナリオの症例に関しては，70％と 90％の違いによって管理戦略が変わってしまうことはないかもしれないが，必ずしもそうとはかぎらない．認知症の検査前確率が 50％の高齢男性で，驚くべきことに SIS で 1 つの誤回答もない第 3 の患者を想定してみよう．2 値化された LR＋/LR－のアプローチを使用した場合（あるいはさらに言えば，感度 / 特異度によるアプローチを使用した場合，数学的に同義で，なおかつ代用可能なため），検査前確率 50％と LR－0.21 の組み合わせから，検査後確率は約 20％となり，さらなる神経心理学的検査やその他の検査が必要となるだろう．表 16–3 からスコア 0 に関連する LR (0.06) を適用した場合，この男性の真の検査後確率はわずか 5％である．この検査後確率であれば，あなた（そして患者およびその家族）は安心して，さらなる検査やさらなる苦痛を免れることができるだろう．

まとめると，多数のカットオフ値や閾値（多重 LR または層別 LR と称されることがある）の使用には，感度 / 特異度アプローチと比べ，2 つの重要な利点がある．第 1 に，連続スコアや多くのカテゴリを示す検査の場合（医療においては多くの検査がそうである），多数の閾値を使用することでできるかぎり多くの情報を保持することができる．第 2 に，ある特定の検査結果の LR を知ると，簡単なノモグラムを使って，自身の患者で，検査前確率から検査後確率を知ることができる．

ユーザーズガイドの適用

ここまで，研究に含まれた患者については結果がおそらく真であることを検証し，検査で取得しうる各スコア別の多重 LRs を計算した．また，どのようにわれわれの患者に結果を適用できるかを示した（ただし，まだ患者のスコアは把握できておらず，スコアが把握できた場合にとるべき措置についても未定である）．

結果を患者のケアにどのように適用できるか

検査結果の再現性とその解釈は，自身の臨床セッティングにおいて満足のいくものか

あらゆる検査の価値は，安定した患者に再適用されたときに同じ結果を出す能力によって決まる．再現性不良は，検査そのもの（例：ホルモン値を決めるための放射免疫測定 radioimmunoassay キットに含まれる試薬のばらつき），または解釈（例：心電図の ST 上昇の程度）に問題があると考えられる．このことは，同じ心電図，超音波検査結果，CT スキャンを吟味しているのに，あなたと一人以上の同僚との間で，（全員が専門家であったとしても）臨床的不一致が浮上する場合を思い浮かべることで容易に確認できる．

理想的には，診断検査に関する論文は，特に解釈の問題に関して，偶然による一致を修正する測定方法を用いて検査結果の再現性を検討するだろう（第 17.3 章「偶然以上の一致の測定」を参照）．

研究セッティングにおいて報告された検査の再現性が中程度で観察者間の不一致がよく起こるものの，その検査が標的疾患の有無をよく区別できる場合には，検査は有用である可能性が高い．このような状況下では，その検査を自身の臨床セッティングへすぐに適用できる可能性が高い．

一方で，もし診断検査の再現性が高いなら，その検査が単純明瞭であるか，またはその検査を解釈する人が熟練しているかのどちらかである．後者の場合には，あなた自身の臨床セッティングで解釈する人が熟練度が低い場合には同じようにはできないかもしれない．適切な研修をする（あるいは自身のセッティングで検査の解釈を行う人がその研修を受けることを確実にする），あるいはより簡単かつより頑強な検査を探す必要がある．

研究結果は自身が診察する患者に適用可能か

検査特性は，疾患重症度の組み合わせや競合疾患の分布が異なることで変わりうる．標的疾患のある患者全員が重症疾患を持つ場合，LRs は 1.0 の値から遠ざかるだろう（つまり感度が増加する）．逆に，患者が全員軽症である場合，LRs は 1.0 の値に近づく（つまり感度が下がる）．標的疾患を持たないが競合疾患を持つ患者の検査結果が，標的疾患を持つ患者の検査結果に類似する場合，LRs は 1.0 に近づき，検査の有用性が低い印象となる（つまり特異度が下がる）．異なる臨床セッティングで，疾患を持たない患者でこのような競合疾患を有する患者が少ない場合は，LRs は 1.0 から離れ，検査の有用性が高い印象となる（つまり特異度が上がる）．

冠動脈疾患の診断において，運動負荷心電図検査により定義される部分母集団によって検査特性が異なる現象が実証されている．血管造影による冠動脈狭窄所見による参照基準とした場合，冠動脈病変が広範であるほど，運動負荷心電図検査での異常の LRs が大きかった[15]．もう 1 つの例は，静脈血栓塞栓症の診断に関するもので，近位静脈血栓症診断のための圧迫超音波検査は，症状の

ない術後患者よりも，症状のある外来患者においてより正確であった [16].

時として，最も診断を必要とする患者において検査が機能しない場合がある．尿路感染症の迅速な診断のための試験紙法検査における陰性結果の LRs は，明らかな症状を呈し，尿路感染症の確率が非常に高い患者では約 0.2 であるのに対し，尿路感染症の確率の低い患者では 0.5 であることから [17]，後者の場合，尿路感染症の可能性を排除するのにほとんど役に立たない．

自身の診療セッティングが研究のものと類似し，懸案の患者が研究の適格基準をすべて満たす場合は，あなたは研究結果が適用可能であるという確信が持てる．そうでない場合は，あなたは判断を下さなければならない．治療介入の場合と同様，自身の患者では疾患重症度や競合疾患の組み合わせがあまりにも異なるために一般化が妥当でないなど，研究結果を自身が診察する患者に適用すべきでない納得のいく理由があるかどうか自問してみるべきである．多くの研究結果を要約するレビューを見つけることができるなら，一般化可能性の問題を解決できるかもしれない [18].

検査結果は自身の管理戦略を変えるか

管理に関わる決断やその伝達に際しては，それを標的疾患の確率と明確に結びつけると有用である．どのような標的疾患にもある確率を下回った時点で臨床医は診断を断念し，それ以上の検査をする必要がないと判断する確率があり，これが検査閾値である．同様に，ある確率を上回った時点で臨床医は診断が確定したとみなし，検査を終了して治療を開始する確率があり，これが治療閾値である．標的疾患の確率が検査閾値と治療閾値の間にある場合は，さらなる検査が必要となる（第 14 章「診断の過程」を参照）.

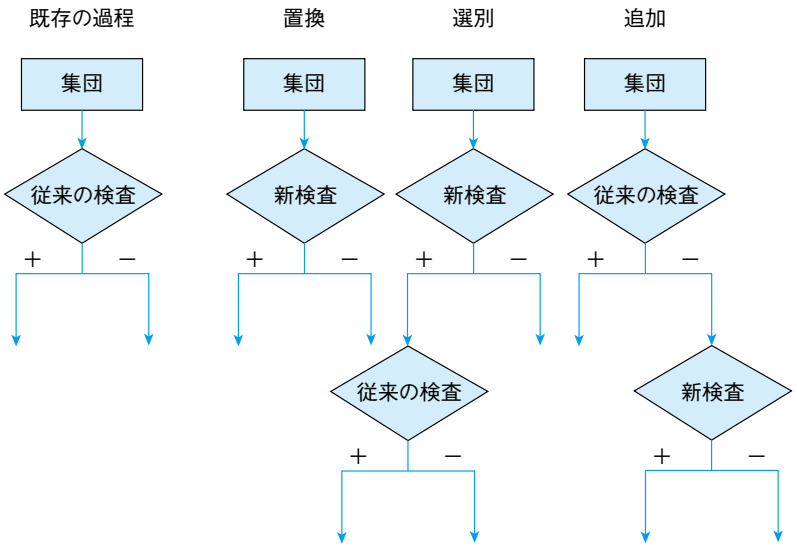
ほとんどの患者の検査結果が 1.0 に近い LRs ならば，検査結果によって検査閾値や治療閾値を超えることは稀であろう．このように，診断検査の有用性は，標的疾患があると疑われる患者群でその検査結果の LRs が非常に高い者，あるいは非常に低い者の占める割合によって大きく影響される．表 16–3 を再検討すると，認知症の疑いのある患者のうち，極端な結果（LR ＞ 10 または LR ＜ 0.1）を持つ者の割合を判断することができる．その割合を計算すると，(105＋2＋64＋2＋11＋163)/(345＋306), すなわち 347/651 ＝ 53%となる．SIS は，認知症の疑いのために検査を受けた患者の半数において検査後確率を確定的に推移させると考えられる．これは注目すべき割合であり，大部分の診断検査よりも優れた結果である.

最後に，逐次検査 sequential tests の使用についてコメントしたい．新検査は，置換 (replacement), 選別 (triage), または追加 (add-on) という 3 通りの方法で既存の診断過程に組み込むことができる（図 16–2）．つまり，既存の診断過程において新検査が既存の検査に取って代わる場合と，従来の検査の前に実施することで，選別のための検査である特定の結果を示した患者にのみ検査過程を続行する場合と，従来の検査の後に位置づけることで，従来の検査にてある特定の結果を示した患者にのみ追加的新検査が必要となる場合がある [19].

診断経路を考慮する上で，LR アプローチは特に適合性が良い．病歴の各項目や身体診察の各所見が診断検査に相当する．われわれは 1 つの検査を使ってある特定の検査後確率を取得でき，その確率はその後実施される別の検査によってさらに増減する．一般に，臨床検査や画像診断も同じ方

図 16–2

既存の診断過程における新検査の 3 つの役割



BMJ Publishing Group の許可を得て，Bossuyt ら[19] より転載.

法で活用することができる．しかしながら，もし 2 つの検査が密接に関係するならば，第 2 の検査を適用したとしてもほとんど，あるいは全く情報は得られないだろうし，LRs の逐次適用によって誤った結果が示されるだろう．たとえば，鉄欠乏の最も強力な臨床検査である血清フェリチンの結果をいったん得ると，血清鉄やトランスフェリン飽和率のような追加的検査はさらに有用な情報をもたらすことはない[20].

臨床予測規則は，一連の検査における独立性の欠如に対処するものであり，それらの結果を統合する方法を臨床医に提供する（第 17.4 章「臨床予測規則」を参照）．たとえば，肺塞栓症が疑われる患者の場合，呼吸症状，心拍数，下肢の症状，酸素飽和度，心電図所見，ならびに病歴や身体診察に関わるその他の局面を組み込んだ規則を活用し，肺塞栓症が疑われる患者を，「高い確率」，「中程度の確率」，「低い確率」に正確に分類することができる[21].

検査結果は患者に利益をもたらすか

診断検査の有用性を判断する最終的な基準は，患者にもたらされる利益が関連するリスクよりも大きいかどうかである[22]．診断検査を適用する利益とリスクを確証するにはどうすればよいのか．その答えは，1 つの治療ツールとして診断検査を考えることにかかっている（第 6 章「治療」を参照）．検査の利益が害よりも大きいかどうかを実証するには，調査中の検査を含む診断戦略に患者をランダム割り付けし，診断戦略または当該検査を伴わない戦略に関連付けられた管理スケジュールを作成し，患者にとって重要なアウトカムの頻度を決めるために両群の患者を追跡する必要がある.

どのような場合に検査の正確さを確認するだけで十分で，どのような場合にランダム化比較試験が

必要となるだろうか．正確な検査の価値は，標的疾患が診断せずに放置されると危険で，検査のリスクが許容範囲で，効果的な治療が存在する場合は議論の余地はない．肺塞栓症の疑いに対する換気血流スキャンの場合がこれに相当する．換気血流スキャンで「高い可能性」，もしくは「正常 / ほぼ正常」という結果は，さらなる検査の必要性を排除し，抗凝固薬が適切に投与される，あるいは適切に差し控えられる（いずれの行動も患者のアウトカムに実質的なプラスの影響を与える）結果となるかもしれない．

完全に無害で，少ない資源投資で，明らかに正確で，明らかに管理に有益な変化をもたらすような検査もある．認知症が疑われる患者における SIS の使用がこれに相当し，検査結果によって認知症でないという確信が得られるか，もしくはさらなる検査の実施が決まり，最終的には症状悪化に対応するための計画が立てられる．

他の臨床状況下では，検査は正確で，その適用によって管理に変化がもたらされるかもしれないが，患者アウトカムへの影響についてははるかに確信性に欠ける場合がある．クリニカル・クエスチョンの定式化についての考察の中でわれわれが提起した問題の 1 つについて考えてみよう（第 3 章「疑問は何か」を参照）．その中で，われわれは肺に明らかに切除可能な非小細胞肺癌を持つ患者について検討し，臨床医はまず CT の実施を指示し，その後の管理については CT の結果を待つべきか，それとも即座に縦隔鏡検査法を実施すべきかについて考えた．この疑問の場合，CT スキャンの正確さを把握しているだけでは不十分である．CT に基づく管理と，全患者を対象とした縦隔鏡検査法とを比較したランダム化試験の実施が必要であり，実際にそのような試験が実施されている [23]．ほかにも，血行動態が定かでない重症患者に対する右心カテーテル検査と，肺感染症の疑いのある重症患者に対する気管支肺胞洗浄などの例がある．これらの検査については，ランダム化試験が最適な管理戦略の究明に役立っている．

ユーザーズガイドの適用

その研究自体再現性を報告していないが，6 つの質問に対する誤回答の数を数えればよいだけなので，採点法は単純でわかりやすい．検査は小道具や視覚的ヒントを必要としないことから，面倒でなく，実施しやすい．SIS の記入には 1 から 2 分しかかからない（のに対し，MMSE の記入には 5 から 10 分かかる）．出版済み論文の付録で，SIS の実施方法が詳細かつ逐一示されていた．あなたは，自分自身もこの検査を確実に実行できると感じた．

臨床シナリオの患者は，一人で通院が可能な高齢の女性だが，以前ほどの頭の冴えはなくなっているようである．本章でこれまで吟味してきた SIS に関する研究における Alzheimer Disease Center のコホートは，介護者が認知症の疑いがあると考え，直接三次医療施設につれてこられた患者らによって構成される．これらの患者の検査特性は一般集団コホート，すなわちそれほど深刻な症状を呈していないサンプルのものと類似すると報告されている．以上からあなたは，研究結果が自身の患者に適用できないとする納得のいく理由はないと判断した．

あなたは，経過観察受診のために患者を診察室に呼び入れ，SIS を実施した．その結果，スコアは 4 であり，検査前確率が 20%だったことから，確率は 60%よりも高い確率へと増加した．患者に対し，患者の記憶力ともしかしたら認知機能にも問題があるかもしれないことを伝えると，患者はさらなる検査を受けるために老年科専門医の紹介を受けることに同意した．

参考文献

1. Callahan CM, Unverzagt FW, Hui SL, Perkins AJ, Hendrie HC. Six-Item Screener to identify cognitive impairment among potential subjects for clinical research. *Med Care*. 2002; 40 (9): 771–781.
2. Thomson DM, Krupey J, Freedman SO, Gold P. The radioimmunoassay of circulating carcinoembryonic antigen of the human digestive system. *Proc Natl Acad Sci U S A*. 1969; 64 (1): 161–167.
3. Lijmer JG, Mol BW, Heisterkamp S, et al. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA*. 1999; 282 (11): 1061–1066.
4. Rutjes AW, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PM. Evidence of bias and variation in diagnostic accuracy studies. *CMAJ*. 2006; 174 (4): 469–476.
5. Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. *Ann Intern Med*. 2004; 140 (3): 189–202.
6. Marantz PR, Kaplan MC, Alderman MH. Clinical diagnosis of congestive heart failure in patients with acute dyspnea. *Chest*. 1990; 97 (4): 776–781.
7. Chochinov HM, Wilson KG, Enns M, Lander S. Are you depressed? screening for depression in the terminally ill. *Am J Psychiatry*. 1997; 154 (5): 674–676.
8. Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics*. 1983; 39 (1): 207–215.
9. Gray R, Begg CB, Greenes RA. Construction of receiver operating characteristic curves when disease verification is subject to selection bias. *Med Decis Making*. 1984; 4 (2): 151–164.
10. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med*. 1978; 299 (17): 926–930.
11. Choi BC. Sensitivity and specificity of a single diagnostic test in the presence of work-up bias. *J Clin Epidemiol*. 1992; 45 (6): 581–586.
12. PIOPED Investigators. Value of the ventilation/perfusion scan in acute pulmonary embolism: results of the Prospective Investigation of Pulmonary Embolism Diagnosis (PIOPED). *JAMA*. 1990; 263 (20): 2753–2759.
13. Fagan TJ. Letter: nomogram for Bayes theorem. *N Engl J Med*. 1975; 293 (5): 257.
14. Sackett DL, Rennie D. The science of the art of the clinical examination. *JAMA*. 1992; 267 (19): 2650–2652.
15. Hlatky MA, Pryor DB, Harrell FE Jr, Califf RM, Mark DB, Rosati RA. Factors affecting sensitivity and specificity of exercise electrocardiography: multivariable analysis. *Am J Med*. 1984; 77 (1): 64–71.
16. Ginsberg JS, Caco CC, Brill-Edwards PA, et al. Venous thrombosis in patients who have undergone major hip or knee surgery: detection with compression US and impedance plethysmography. *Radiology*. 1991; 181 (3): 651–654.
17. Lachs MS, Nachamkin I, Edelstein PH, Goldman J, Feinstein AR, Schwartz JS. Spectrum bias in the evaluation of diagnostic tests: lessons from the rapid dipstick test for urinary tract infection. *Ann Intern Med*. 1992; 117 (2): 135–140.
18. Irwig L, Tosteson AN, Gatsonis C, et al. Guidelines for meta-analyses evaluating diagnostic tests. *Ann Intern Med*. 1994; 120 (8): 667–676.
19. Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. *BMJ*. 2006; 332 (7549): 1089–1092.

20. Guyatt GH, Oxman AD, Ali M, Willan A, McIlroy W, Patterson C. Laboratory diagnosis of iron-deficiency anemia: an overview. *J Gen Intern Med*. 1992; 7 (2): 145–153.

21. Wells PS, Ginsberg JS, Anderson DR, et al. Use of a clinical model for safe management of patients with suspected pulmonary embolism. *Ann Intern Med*. 1998; 129 (12): 997–1005.

22. Guyatt GH, Tugwell PX, Feeny DH, Haynes RB, Drummond M. A framework for clinical evaluation of diagnostic technologies. *CMAJ*. 1986; 134 (6): 587–594.

23. Canadian Lung Oncology Group. Investigation for mediastinal disease in patients with apparently operable lung cancer. *Ann Thorac Surg*. 1995; 60 (5): 1382–1389.