特別寄稿 5
診断に関する診療ガイドライン（CPG）の作成

森實敏夫[1]、河合富士美[2]、小島原典子[3]

[1] 日本医療機能評価機構、[2] 聖路加国際大学学術情報センター図書館、
[3] 東京女子医科大学衛生学公衆衛生学第二講座

2015 年 12 月 15 日　掲載

0 はじめに

診断法に関する診療ガイドライン作成の手順は治療法に関する診療ガイドライン作成手順と同じである。

1. 臨床的文脈の中で取り上げるべき臨床課題を決める。
2. 臨床課題に基づきクリニカルクエスチョンを作成する。

クリニカルクエスチョンごとに：

3. 益のアウトカムと害のアウトカムをリストアップして重要性を決める。
4. エビデンスを収集する。
5. アウトカムごとにエビデンスを評価する。
6. アウトカムごとにエビデンスを統合し（システマティックレビュー）エビデンス総体の強さを評価する。
7. エビデンスの強さ（効果の大きさと不確実性）、益、不利益（害、負担、費用）、患者・介護者の価値観や好みを評価して推奨の強さを決める。

しかしながら、臨床的文脈における位置づけ、アウトカムの患者中心性、文献検索法、エビデンス評価の項目、効果指標、メタアナリシスの手法の点で異なっている。以下、主に治療に関する診療ガイドライン作成と異なる点について述べる。

1　診断の CPG の特徴

1.1　診断法研究のレベル

診断法に 関しては、システマティックレビュー（ＳＲ）、推奨作成の方法において、明示的で透明性の高い枠組みを適用している例は少ない[1]。GRADE ワーキンググループでも診断法の SR 作成の方法論が検討されているが[2, 3, 4, 5]、治療の取り組みと比較しても診断の CPG 方法論の確立は遅れていると言えよう[6]。

診断法研究のレベルとしては以下の６つが考えられている[7]。

表１．診断法研究のレベル

レベル	効果	研究目的
1	技術の確立と最適化	安定した結果
2	診断能（精度/正診率）	感度・特異度
3	診断思考への影響	検査実施後医師の疾患確率評価が変わる率
4	治療選択への影響	検査実施後に治療計画が変更される率
5	患者アウトカムへの影響	検査を実施しない場合と比較して実施した場合にアウトカムが改善する（生存、QOL など）
6	社会への影響	費用効果分析（例：検診における有用性）

表 1 の 5 に該当する患者にとって重要なアウトカムに注目した診断研究がもっとも重要と考えられるが、エビデンスを統合できる研究はスクリーニング法などの一部の診断法に限られる。現状では、診断精度[a]に関する横断研究が多く、複数の研究から診断精度の指標である感度・特異度を統合するメタアナリシスが行われている。

1.2　診断の種類

方法論としてまとめられている多くの先行研究においては、Diagnostic test と表記され、患者の臨床診断、集団に対するスクリーニング、サーベイランスの 3 つに大別されている。また、一次的な診断だけでなく、治療応答の評価や、Interventional radiology の際に用いられる診断法の利用もある。本稿では、臨床面接から得られる症状、身体所見をふくめた臨床診断の、検査、一連の検査、一組の診断手順を参照基準と比較する、臨床診断の CPG の作成方法について記載する。

[a] Accuracy は精度または正確度、Diagnostic test accuracy（DTA）は診断精度とする。対象者の事前確率による的中率 Predictive value の変動も含めた診断能 Diagnostic performance は正診率とする。

1.3 診断精度の指標

（１）感度 sensitivity と特異度 specificity

感度・特異度は診断法固有の属性であるが、診断閾値によって変動し、疾患スペクトルの影響を受ける。正診率は 事前確率によって変動するため、プライマリケア、二次・三次ケアで異なり、スクリーニング・サーベイランス・診断などの用途で異なる。

（２）尤度比 LR, Likelihood ratio　　感度・特異度から導出される指標

（３）診断オッズ比　DOR, Diagnostic odds ratio　　感度・特異度から導出される指標

これらの指標は真陽性（TP）、偽陰性（FN）、真陰性（TN）、偽陽性（FP）（表 4 参照）の人数から算出され、感度＝TP/(TP+FN)、特異度＝TN/(TN+FP)、陽性尤度比=[TP*(TN+FP)]/[(TP+FN)*FP]、陰性尤度比=[FN*(TN+FP)]/[(TP+FN)*TN]、診断オッズ比＝TP*TN/FN*FP で算出される。

2 作成方法

2.1 作成グループの選出

治療の CPG 作成と同様、学際的なメンバーの選出と利益相反の公開が必要であり、特定の利益相反のあるメンバーは、関連する推奨の決定に加わらないなどの配慮が求められる。特に、作成委員長は利益相反が問題にならないメンバーが選出されるべきである。作成グループ全体で、予め、ガイドラインが必要な重要な臨床課題を抽出し、作成方法を決定する。

2.2 クリニカルクエスチョンの定型化

診断法に関するクリニカルクエスチョンの類型は以下の５つが考えられる。

（１）いずれの診断法の精度が高いかを問う。
　　例：疾患 X が疑われる場合どの診断法を選択すべきか?
（２）疾患確率が変わる結果が得られるかを問う。
　　例：医師の鑑別診断の順序が結果により変化するか?
（３）治療法の選択が変わるかを問う。
　　例:PET スキャンで転移が検出された場合、化学療法が選択される率は上がるか？
（４）患者関連アウトカムの改善を問う。
　　例：2 年おきの乳房撮影で乳癌の死亡率が下がるか？
（５）費用効果分析。
　　例：ｘｘ癌検診によりｘｘ癌の医療費が減少するか？

診断法に関するクリニカルクエスチョンは介入に関する場合と原則的に同じ形式を用いることができる。対象 P は診断標的が疑われる者、介入 I はインデックス診断法（の実施）、対照 C は比較診断法（の実施）または実施せず、アウトカム O は死亡、QOL、病的状態[b]（診断法の害、治療の害を含む）などの患者中心アウトカムが設定される。また、間接的アウトカムとしては診断法の技術的特性、診断精度、臨床決断への影響、治療選択への影響が設定される。インデックス診断法は診断能を解析する対象である。2 つの診断法を直接比較（Head-to-head の比較）する場合には、インデックス診断法と比較診断法が比較される。アウトカムとして、診断標的が設定される場合は、参照基準がその存在の有無を決定する基準として用いられるので、診断標的（参照基準）というように記述する。

例としては「急性虫垂炎が疑われる成人で腹部超音波検査は急性虫垂炎の診断に有用か？」などがあげられる。この例では、診断標的は“急性虫垂炎（切除虫垂の病理診断）”であり、それをアウトカムに設定している。上記のアウトカムの解説の内、病的状態に相当

[b]診断法で検出しようとする病的状態という意味で、診断標的のことおよび診断法自体に伴う害とその後実施される治療に伴う害を意味する。

する。

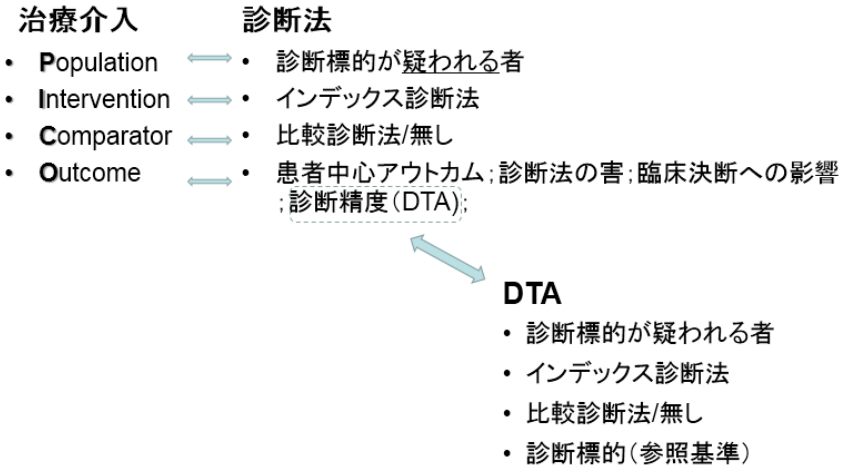


図 1．診断法に関する CQ の構成要素

2.3 文献を系統的に検索する

検索式は原則として、「インデックス検査 （OR 参照基準） AND 診断標的」から構成され、さらに必要に応じて「AND 検索フィルター」が組み合わされる。

それぞれのデータベース毎に多数の診断用検索フィルターが提案されているがシステマティックレビュー作成にはどれも用いるべきでないとの Cochrane Review も報告されている。しかし、検索結果の論文数が非常に多い場合、Number needed to read（NNR）が非常に大きい場合、フィルターの感度と特異度を認識し漏れのリスクと作業効率のバランスを考慮した上でいずれかのフィルターの使用を検討してもよい。

また、すでに診断精度に関するシステマティックレビュー/メタアナリシスが発表されている場合には、非直接性の評価を含めたクリニカルクエスチョンへの適合度、AMSTAR による妥当性評価結果、に基づいて採用を考える。

PubMed の Clinical Queries では診断に関する検索のための下記のフィルター（Sensitive/Broad）が用いられている。その感度は 98%、特異度は 74%と報告されている[8]。

(sensitiv*[Title/Abstract] OR sensitivity and specificity[MeSH Terms] OR diagnos*[Title/Abstract] OR diagnosis[MeSH:noexp] OR diagnostic * [MeSH:noexp] OR diagnosis,differential[MeSH:noexp] OR diagnosis[Subheading:noexp])

1）データベース

＜ガイドライン＞

National Guideline Clearinghouse(NGC)　　http://guideline.gov/

NICE Evidence Search　http://www.nice.org.uk/
International Guideline Library　http://www.g-i-n.net/library/international-guidelines-library　＊
＜文献＞
PubMed/MEDLINE　http://www.ncbi.nlm.nih.gov/pubmed
医中誌 Web　http://www.lib.twmu.ac.jp/protected/ichuindex.html　＊
The Cochrane Library　http://www.thecochranelibrary.com/view/0/index.html　＊
＊要契約

EMBASE や JMEDPlus, CINAHL,PsychInfo®なども必要に応じて追加する[9]。

２）一次スクリーニング

タイトル、アブストラクトから明らかに、主題の異なる研究を振るい落とす一次スクリーニングを行う。2 名のレビュアーにより独立して行う。2 名の結果を照合し、2 次スクリーニング用データセットを作成し、文献を収集する。

３）二次スクリーニング

2 名のレビュアーにより独立して作業し、意見が異なる場合は第 3 者の意見を取り入れ採用論文を決定する。文献の本文より感度・特異度など診断の正確度の指標が数値化されているものを選択する。

2.5 システマティックレビュー

１）個別研究の評価

診断精度研究の論文執筆ガイダンスである STARD (Standards for the Reporting of Diagnostic accuracy studies)[10] や診断精度研究の質評価のチェックリストである QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies-2)[11]**【参考資料１】**、などを研究評価に用いることができる。コクランの Diagnostic Test Accuracy Working Group の Handbook for Diagnostic Test Accuracy Reviews[12] の第 9 章 Assessing methodological quality も参考になるが、ハンドブックは未完成の状態である。

これらを参考に、個別診断精度研究および診断精度エビデンス総体の評価シートを作成した**【テンプレート１】**。

診断精度研究の研究デザインは横断研究であるが、いわゆる症例対照研究型の研究とコホート研究型の研究がある。前者では、診断しようとする標的疾患を有する集団と、それ以外の集団で診断法を実施し、その感度・特異度を解析する研究手法が用いられたものであり、Two-gate study と呼ばれる。このタイプの研究では対象者が実際の臨床とは異なっているため、そのまま臨床に適用することが困難となる。一方、後者では、ある一定の症状や検査結果を呈した集団で参照基準となる診断法および解析対象であるインデックス診断法を実施し感度・特異度を解析する研究手法であり、Single-gate study と呼ばれる。このタイプ

の研究は対象者が実際の臨床と同じである場合が多いので、そのまま臨床に適用することが可能である。研究デザインの判定時には RCT、コホート研究、症例対照研究、横断研究などの分類だけでなく、Two-gate study か Single-gate study かを明らかにする必要がある。通常、メタアナリシスによって感度・特異度の統合値を算出する場合には Single-gate study だけを対象とする。

診断法の研究のバイアスリスクのドメインおよび（評価項目）は、選択バイアス（臨床に即したランダム選択）、インデックス検査（盲検化）、参照基準（盲検化、不完全な参照基準）、症例減少バイアス（不完全な検査実施）、フローとタイミング（同時期に実施）、その他（データ欠損など）である。個別研究について、これらは”高(-2)”、”中/疑い(-1)”、”低(0)”の 3 段階で評価する。

非直接性はクリニカルクエスチョンの対象、インデックス検査、参照基準、アウトカムと各研究のこれら項目との一致性を評価する。上記のように 3 段階評価を行う。

QUADAS-2 とテンプレート 1 で示す評価シートではいくつかの異なる点があるが、評価項目として同じ概念のものを全て含んでいる。相違点については**【参考資料 3】**にまとめた。

感度・特異度の算出の基になった TP,FP,FN,TN の人数を記述し、研究対象の有病率（事前確率）とその信頼区間、感度・特異度とその信頼区間、正診率とその信頼区間、ROC（Receiver operating characteristic）解析が行われている場合は AUC（Area under the curve）とその信頼区間を抽出し、記述する。TP,FP,FN,TN の人数はメタアナリシスで感度・特異度の統合値を算出する際に用いられる。

2）エビデンス総体の評価

エビデンス総体ではバイアスリスク、不精確、非一貫性、不精確、非直接性、その他（出版バイアスなど）のドメインについて同じく”高(-2)”、”中/疑い(-1)”、”低(0)”の 3 段階で評価する。バイアスリスクは個別研究の評価結果のまとめの部分から複数の研究の全体としてのまとめとして評価する。非直接性についても同様である**【テンプレート２】**

疾患があると正しく分類された対象者の人数 TP、疾患がないと正しく分類された対象者の人数 TN、疾患があると誤って分類された対象者の人数 FP、疾患がないと誤って分類された対象者の人数 FN を対象者が 1000 人の場合について算出し、その診断法の実施によってどれくらいの益をどれくらいの人が得ることができるか、どれくらいの害をどれくらいの人が被るのかを評価する際に参考とする。

さらに、診断精度研究における患者にとって重要なアウトカムの重要度を 0～9 で数値化し、各アウトカムについて患者にとっての重要度を判定する。

3）益と害の評価

診断法の益は疾患に罹患していることが正しく診断された結果、すなわち TP、その疾患に対応した治療を受けることによってもたらされる。また、疾患に罹患していないことが正しく診断された結果、すなわち TN、特に治療を受けないで済むことによる益もある。一方

で、診断法の害は診断法の実施に伴って起きる直接的な害、たとえば大腸内視鏡で起きる腸管穿孔のようなものと、正しく診断されなかった結果受ける害がある。正しく診断されなかった場合の内 FP の場合は、不必要な治療を受けることが害を構成する。FN の場合は、必要な治療を受ける機会を失うことで、本来治療によってもたらされるはずの益を受けられなくなるという害が生じる。

したがって、診断法の益と害は、診断法の負担や直接の害だけでなく、その後の治療法の益と害とを一緒に考えないと評価できない。治療法の益は疾患のアウトカムと治療効果の大きさによって決まるので、同じ感度・特異度の診断法であっても、対象疾患によってその意義は変わってくる。

多くの場合は、全体としてどれくらいの益をどれくらいの人が受けるか、どれくらいの害をどれくらいの人が受けるかを常識的に評価することで益と害のバランスを判定することが可能であろう。

より定量的に評価するためには以下に述べる決定木 Decision tree を用いるのが一つの方法である [13]。決定木を用いる場合には、それぞれの選択肢の価値を効用という概念で定量化する必要があり、それが容易でないという課題がある。基準的賭け法 Standard gamble などを用いることが可能であるが、詳細は省略する。

疾患に罹患していることを D+、罹患していないことを D-、治療を行うことを A+、行わないことを A-で表し、効用 Utility を U で表すと、疾患に罹患していて治療を受ける効用は U(D+A+)で表される。同様に、U(D-A+)は疾患で無いのに治療を受ける効用、U(D+A-)は疾患なのに治療を受けない効用、U(D-A-)は疾患でなく治療も受けない効用を表す。通常、U(D-A-)>U(D+A+)>U(D-A+)>U(D+A-)の順で効用が大きいと考えられる。

上記の表記を用い、診断法が陽性を T+、陰性を T-で表して決定木を作成すると図 2 のようになる。

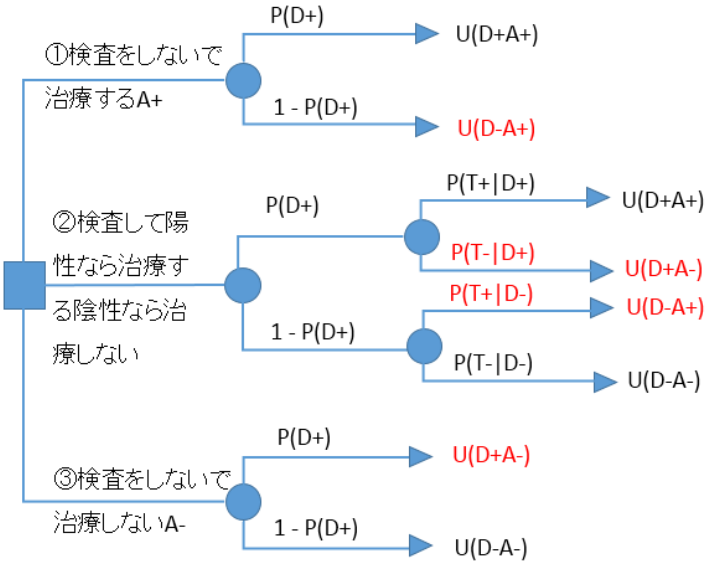


図 2. 検査（診断法）を実施する効用と決定木

この決定木によって、診断法を実施するという選択肢②の効用は:

[U(D+A+)*P(T+|D+)+U(D+A−)*P(T−|D+)]*P(D+)+[U(D−A+)*P(T+|D−)+U(D−A−)*P(T−|D−)]*(1−P(D+))+U(T)

となる。最後の項 U(T)は診断法に伴う害、負担、費用、患者の意向を総合した効用値である。通常は負の値になる。この選択肢②の効用値を、診断法を実施しないで治療する選択肢①および診断法を実施しないで治療をしない選択肢③の効用値と比べて、いずれよりも大きければ診断法を実施すべきであるという結論が得られる。

P(T+|D+)感度 Se であり、P(T−|D+)は１－Se、P(T−|D−)は特異度 Sp であり、P(T+|D−)は１－Sp に相当し、P(D+)は疾患確率であり、これを Pr で表し、式をこの表記で置き換えると診断法を実施する効用は次のようになる：

[U(D+A+)*Se+U(D+A−)*(1−Se)]*Pr+[U(D−A+)*(1−Sp)+U(D−A−)*Sp*(1−Pr)+U(T)

感度 Se、特異度 Sp は診断法固有の属性であるが、その実施の効用は有病率、適用される治療法の益と害によって決まり、感度・特異度だけでは決まらないことがわかる。

また、この決定木で用いられる各選択肢の効用の評価には治療法に伴う益と害の評価が包含される。たとえば、U(D+A+)を決める際に、その効果が高くても副作用がひどければ、低く見積もることになり、同じ程度の効果で副作用がほとんど無ければ、高く見積もることになり、益だけで決めるわけではない。

図 2 に示す U(D+A+)と U(D+A−)の差、すなわち U(D+A+)−U(D+A−)は治療によってもたらされる益に相当する。また、U(D−A−)と U(D−A+)の差、すなわち(D−A−)−U(D−A+)は費用、負担、害すなわち、不利益に相当する。

従来、益を B、不利益を C で表すと、治療閾値、すなわち治療を受けることの効用が治療を受けないことの効用を上回る疾患確率は C/(B+C)となることが知られている。それは、図 2 の選択肢①の効用は Pr*U(D+A+)+(1−Pr)*U(D−A+)、選択肢③の効用は Pr*U(D+A−)+(1−Pr)*U(D−A−)となりそれぞれの終末の効用が一定であれば、これらの効用は疾患確率 Pr の関数となり、Pr とこれらの 2 つの選択肢の効用は直線関係にあることから得られる結果である。図 3 にこれらの関係を示す。

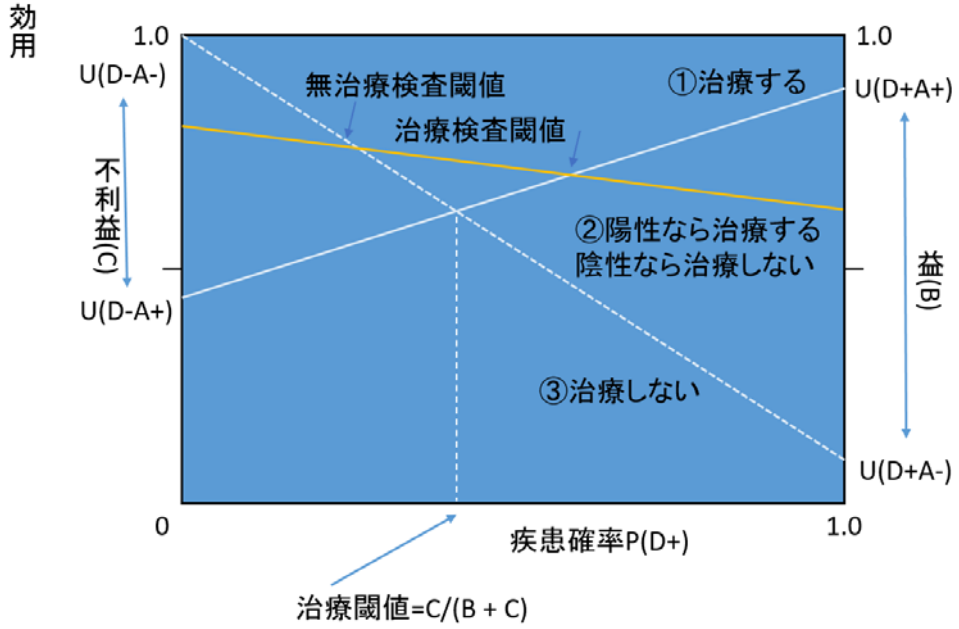


図 3. 疾患確率と 3 つの選択肢の効用の関係。

無治療検査閾値は検査をしないで治療をしない効用と検査をして治療をしない効用が同じになる疾患確率である。無治療検査閾値を超える疾患確率の場合は、検査を行って、陽性なら治療し陰性なら治療しない選択肢の方が効用値が高い。治療検査閾値は検査をしないで治療をする効用と検査をして治療をする効用が同じになる疾患確率である。治療検査閾値より低い疾患確率の場合は、検査を行って、陽性なら治療し陰性なら治療しない選択肢の方が効用値が高い。

それぞれの値は、感度・特異度と益 B と不利益 C によって決定され、以下の式で算出される [13]。上記の治療閾値=C/(B+C)に対して、診断法の診断能の指標である感度 Se と特異度 Sp がかかわっていることがわかる。疾患確率が、無治療検査閾値と治療検査閾値の間にある場合は、検査を実施すべきであり、陽性の結果が得られれば疾患確率が治療閾値を上回り、陰性の結果が得られれば治療閾値を下回ることになる。

無治療検査閾値=[(1 − Sp)・C]/[(1 − Sp)・C + Se・B]
治療検査閾値=[Sp・C/[Sp・C+(1 − Se)・B]

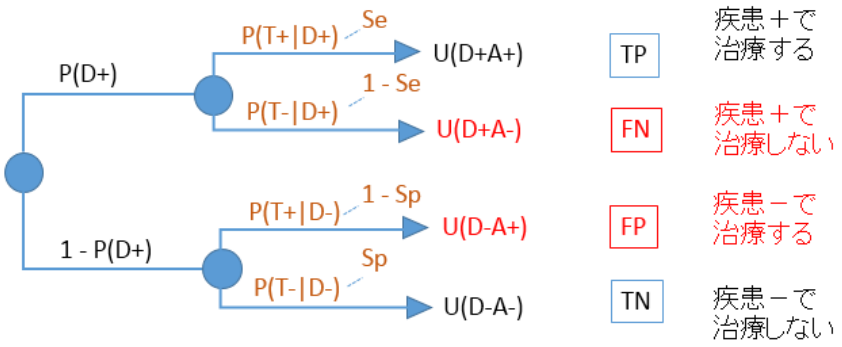
さらに、診断法に伴う害、負担、費用、患者の意向を総合した効用値である上記の U(T)を不利益に加算する場合は以下の式で算出される。

無治療検査閾値=[(1 − Sp)・C − U(T)]/[(1 − Sp)・C + Se・B]
治療検査閾値=[Sp・C − U(T)]/[Sp・C+(1 − Se)・B]

診断法実施の効用には有病率あるいは事前確率も影響することから、その診断法の適用を想定する対象者の有病率が異なる場合には、結果が変わる可能性があることも認識しておく必要がある。Single-gate study で選択バイアスが無ければその研究対象の有病率に基づいた解析だけで考えても問題が無い。その場合の有病率は TP+FN/(TP+FN+TN+FP)で求められる。

個別患者にその診断法を適用する場合には、事前確率（検査前確率）が個々に異なるので、個別の判断が必要になるが、対象者全体としての効用の判定は研究対象の有病率に基づいた解析から可能である。

診断法を実施する選択肢の部分を取り出し、TP, FN, FP, TN に該当する部分を示したものが図 4 である。



診断法を実施し陽性なら治療し陰性なら治療しない方針の場合の効用と感度 Se・特異度 Sp、事前確率 P(D+)の関係を示す決定木。赤字で示す部分が害を生成しうる部分である。

図 4. 診断法実施の結果に基づいて治療方針を決定する場合の決定木

多くの場合は、全体としてどれくらいの益をどれくらいの人が受けるか、どれくらいの害をどれくらいの人が受けるかを評価する際にはこのような決定木を念頭においておくとよい。

また、結果が連続変数で表される検査のようにカットオフの設定によって感度・特異度が変動しうる場合がある。その場合に正診率を最大化するカットオフ値の設定が益を最大化し害を最小化することになるとは限らない。たとえば、診断後の治療法の害が大きい場合、感度は低くなっても特異度を大きくするようにカットオフ値を設定した方がその診断法の実施によりもたらされる益がより大きく、害がより小さくなることがありうる。ROC 曲線上でカットオフ値を設定する場合、本当は疾患が無いのに治療を受けること（FP）による正味の害を H とし、本当に疾患があり治療を受けること（TP）による正味の益を B とし、事前確率を P(D)とした場合、傾きが $H\times[1-P(D)]/B\times P(D)$ の ROC 曲線の接線の接点が益を最大化し害を最小化するカットオフ値に対応する。

4）診断法のメタアナリシス

エビデンス総体の評価の際には、メタアナリシスによって得られる感度・特異度の統合値と信頼区間を用いる。

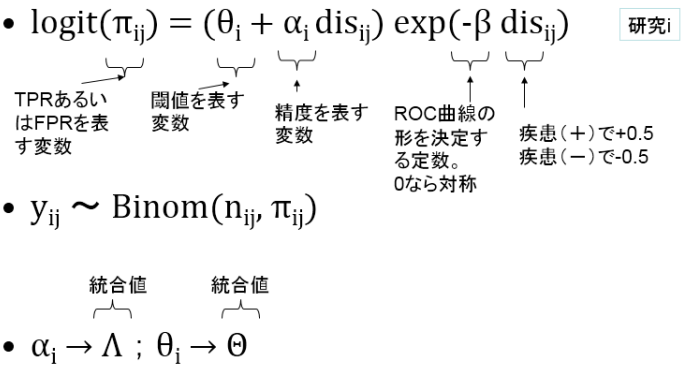
感度および特異度は割合（率）であるから、割合の分散の逆数で重み付けし他の効果指標の場合と同じように、統合値と信頼区間を算出することができる。このようなプール解析ではプールした値が 2 つとそれぞれの信頼区間の値が得られるが、感度・特異度の両者の関係とカットオフ値の変動を考慮した診断能あるいは診断精度を表す指標が得られるわけではない。

感度・特異度をそれぞれオッズに変換し、その比すなわち診断オッズ比を算出しそれを通常のメタアナリシスの手法で統合することも可能であるが、この場合もカットオフ値の問題は無視されることになる。

Moses & Littenberg の Summary ROC はカットオフ値をモデルに取り込んだ方法であるが[14,15]、個々の研究のサンプリングエラーが十分考慮されていない、感度・特異度の平均値が出せないなどの欠点が指摘されている。

Reitsma らは Moses & Littenberg のモデルの欠点を改良した二変量モデル Bivariate model を開発し[16]、Doebler P はそのための R のパッケージ mada を発表している[17]。

Rutter & Gatsonis が階層サマリーROC(Hierarchical SROC, HSROC)の方法を発表し[18]、研究間のバラツキを取り込むことができるモデルを発表した（図 5）。



j は疾患有り無しを表し 1 または 0 の値をとる。i は研究番号を表す。統合値Λと Θおよびその標準誤差から感度・特異度の値が計算される。感度の期待値は E(TPR) = 1/(1 + exp{-[(Θ+Λ/2)exp(-ß/2)]})で、偽陽性率の期待値は E(TNR) = 1/(1 + exp{-[(Θ-Λ/2)exp(ß/2)]})で算出される。

図 5. Rutter & Gatsonis の HSROC モデル

mada を用いた Bivariate model でランダム効果モデルによる解析の一例をスクリプトとともに以下に示す。

```
###meta-analysis with mada using tab-separated tabular data###
library("mada")              #Load mada package
data=read.table("file name.txt", sep="¥t", header=TRUE)
#data=read.delim(" clipboard",sep="¥t",header=TRUE)      #Windows
#data=read.delim(pipe("pbpaste"),sep="¥t",header=TRUE)  #Mac
data.d=madad(data)
forest(data.d, type="sens", snames=data$names, cex=0.65,main="Sensitivity")
dev.new();forest(data.d, type="spec", snames=data$names, cex=0.65,main="Specificity")
fit.reitsma=reitsma(data)
summary(fit.reitsma)
xmax=0.5;ymin=0.2                  #Set max value for x-axis and min value for y-axis.
dev.new();plot(fit.reitsma, sroclwd=2, main="SROC (bivariate model)",
xlim=c(0,xmax),ylim=c(ymin,1))
points(fpr(data),sens(data),pch=2)
legend("bottomright", c("Each study", "Summary estimate"),pch=c(2,1))
legend("bottomleft",c("SROC", "Conf. region"),lwd=c(2,1))
```

解析されるデータは次のようなフォーマットでタブ区切りのテキストファイルとして用意する。一行目がラベルで以下行ごとに研究 ID、TP、FP、FN、TN の人数である。研究 ID は第一著者名＋スペース＋年度である。あるいは Microsoft Excel で用意したデータを、ラベルからデータの最後の行までの範囲を選択しコピー操作をした上で、クリップボード経由で読み込んで解析することも可能であり、その場合は上記の 4 行目または 5 行目のスクリプトを用いる。

```
names TP FP FN TN
Theron E 2013 154 27 31 517
Malbruny TN 2011 12 0 0 46
・・・・・
```

	A	B	C	D	E
1	names	TP	FN	FP	TN
2	Theron 2013	154	31	27	517
3	Malbruny 2011	12	0	0	46
4	Boehme 2011e	101	0	16	671
5	Boehme 2011b	171	6	3	825
6	Boeheme 2010b	201	8	0	101
7	Ciftci 2011	24	1	1	59

図 6. 解析のためのデータ。

感度および特異度は Forest plot として表示することができる（図 7）。なお、Funnel plot が必要な場合は、DOR を指標として通常のオッズ比の Funnel plot を R の metafor パッケ

ージなどを利用して作成することができる。しかし、出版バイアスの検出力には問題があることが指摘されており [19]、参考程度に用いるべきであろう。

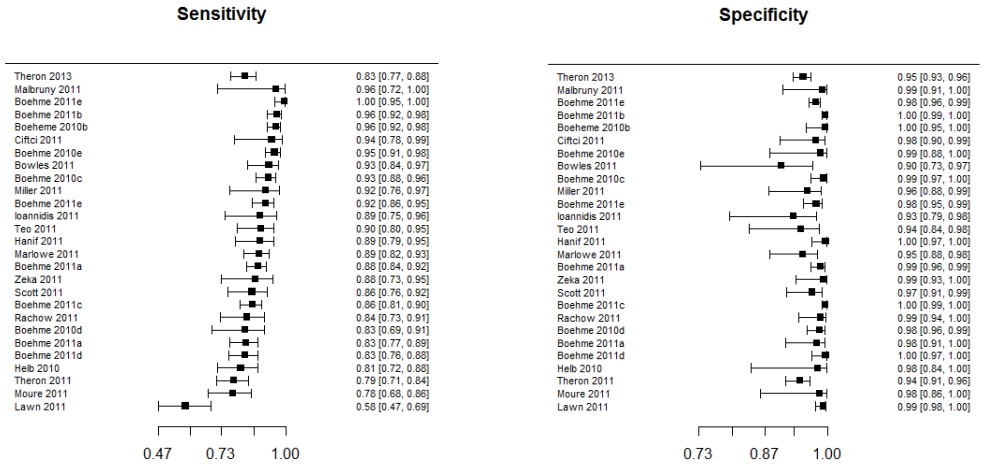
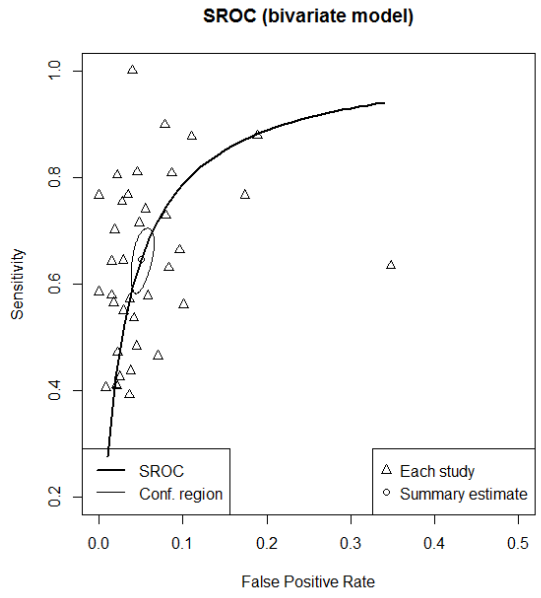


図 7.　感度・特異度の Forest plot

結果として得られた SROC 曲線を図 8 に示す。このモデルは HSROC と同等である。



関節リウマチにおける抗シトルリン化ペプチド抗体の診断能。AUC 0.846、感度 0.66(95%信頼区間 0.592-0.696)、特異度 0.95(0.935-0.961)。個々の△が一つの研究の結果を表し、○が感度・特異度の統合値に相当する点、楕円形がその 95%信頼区間の範囲を表している。左上と右中央の 2 つの研究は外れ値の可能性が高いことがわかる。

図 8.　mada による Bivariate model を用いた SROC の一例

この分野の研究はさらに発展し、参照基準が完全な場合だけでなく、不完全な場合、さらに参照基準が異なる研究を統合する場合、参照基準と評価診断法に相関がある場合などに対応する手法が開発され、解析のためのコードや実例も発表されている。たとえば、Dendukuri N[20] は、論文だけでなく自身のウェブサイト [21] でも解説のスライドや R、WinBUGS、SAS 用のコードを発表している。必要に応じて、これらのベイジアン法による解析を行う。その理論および実際の OpenBUGS あるいは WinBUGS 用のコードは**【参考資料 2】**に示す。

また AHRQ は DTA 研究のメタアナリシスのさまざまな手法を比較した報告書 [22] を発表しており、理論的には二変量モデルと用いることが望ましいこと、また、ベイジアン法では不確実性を完全に定量化でき外部のエビデンスを取り込むことができるため二変量モデルでは十分解析できないような場合でも有用であることを述べている。

2.7　推奨の決定

推奨の決定は治療法の診療ガイドラインの場合と基本的には同じであり、エビデンスの強さ、益と害のバランス、患者の価値観や好み、負担、コストや資源配分を考慮した上で、その強さを決定する。

患者関連アウトカムへの効果を直接評価できない場合が多いので、エビデンスの限界を正しく評価して、患者の益を推定しなければならない。

2.8　外部評価

独立した評価委員会の評価を受ける。評価の結果とそれをどのように反映させたかを記録する。

2.9　フィードバック

スコープの段階と診療ガイドラインの草稿の段階で、対象者、使用者、ステークホルダーからのフィードバックを公式な形で得て、その結果を最終稿に反映させるとともに、変更の過程を記録する。

テンプレート

【テンプレート１ ：個別診断精度研究評価シート】

評価シート＿診断精度研究用

CQ	
対象	
インデックス検査	
対照	
参照基準	

*バイアスリスク、非直接性
各項目の評価は"高(-2)"、"中/疑い(-1)"、"低(0)"の3段階
まとめは"高(-2)"、"中(-1)"、"低(0)"の3段階でエビデンス総体に反映させる

まとめは"高(+2)"、"中(+1)"、"低(0)"の3段階でエビデンス総体に反映させる
各アウトカムごとに別紙にまとめる

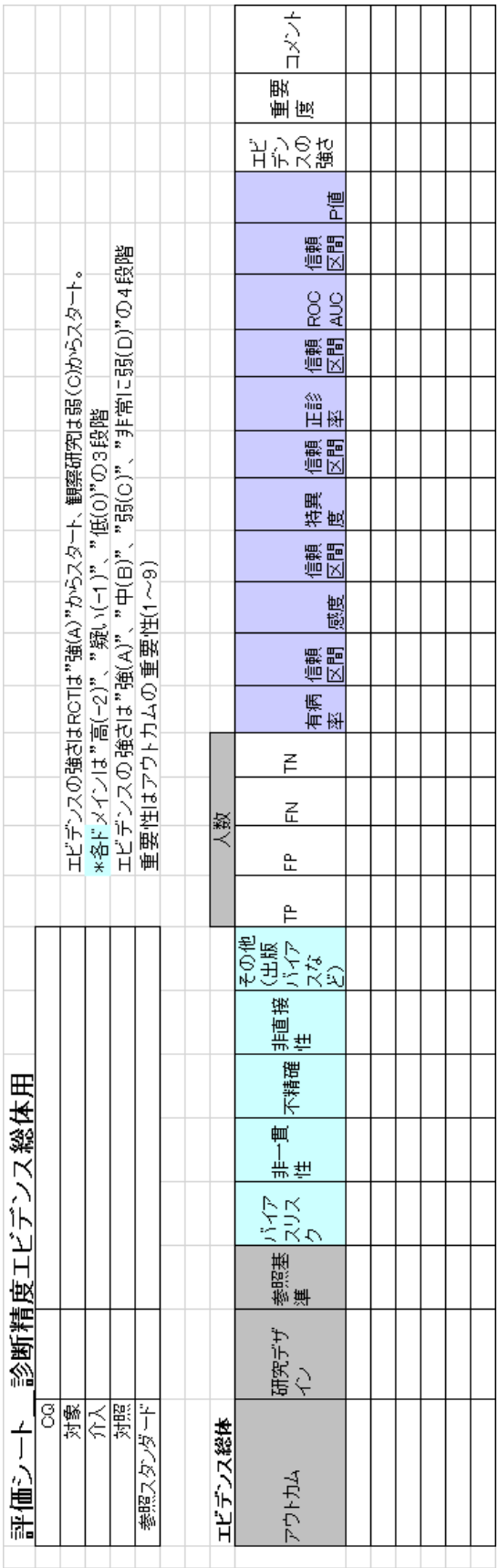
アウトカム	

個別研究

			バイアスリスク*								非直接性*					人数														
			選択バイアス	インデックス検査	参照基準		症例減少バイアス	フローとタイミング	その他																				
研究コード	研究デザイン	参照スタンダード	臨床に即したランダム選択	盲検化	盲検化	不完全な参照基準	不完全な検査実施	同時期に実施	データ欠損など	まとめ	対象	インデックス検査	参照基準	アウトカム	まとめ	TP	FP	FN	TN	有病率	信頼区間	感度	信頼区間	特異度	信頼区間	正診率	信頼区間	ROC AUC	信頼区間	P値

コメント（該当するセルに記入）

【テンプレート２　：診断精度エビデンス総体評価シート】



評価シート　診断精度エビデンス総体用

CQ	
対象	
介入	
対照	
参照スタンダード	

エビデンスの強さはRCTは”強(A)”からスタート、観察研究は弱(C)からスタート。
＊各ドメインは”高(-2)”、”疑い(-1)”、”低(0)”の3段階
エビデンスの強さは”強(A)”、”中(B)”、”弱(C)”、”非常に弱(D)”の4段階
重要性はアウトカムの重要性(1～9)

エビデンス総体

アウトカム	研究デザイン	参照基準	バイアスリスク	非一貫性	不精確	非直接性	その他（出版バイアスなど）	人数 TP	FP	FN	TN	有病率	信頼区間	感度	信頼区間	特異度	信頼区間	正診率	信頼区間	ROC AUC	信頼区間	P値	エビデンスの強さ	重要度	コメント

参考資料

【参考資料１ ：QUADAS-2】

フェーズ１：レビュークエスチョンの記述

患者（セッティング，インデックス検査の使用目的，症状，事前検査）：
インデックス検査：
参照基準および診断標的：

フェーズ２：各レビューのための調整**

フェーズ３：一次研究のフローダイアグラム作成**

フェーズ４：バイアスリスクと適用可能性の判定**

ドメイン１：患者選択 **A. バイアスリスク**
患者選択方法の記述：
・連続した患者かランダムサンプルを組み入れたか。　はい/いいえ/不明 ・症例対照デザインではないか。　はい/いいえ/不明 ・その研究は不適切な除外を行っていないか。　はい/いいえ/不明 **患者選択はバイアスを生じた可能性があるか。　リスク：低　/高　/不明** **B.適用可能性に関する懸念**
組み入れられた患者の記述（事前検査，症状，インデックス検査の使用目的およびセッティング）：

組み入れられた患者はレビュークエスチョンに合致していない懸念があるか。　　**懸念：低　/高　/不明**

ドメイン 2：インデックス検査 **A. バイアスリスク**
インデックス検査と，それがどのように実施され解釈されたか記述：
・インデックス検査の結果の解釈は，参照基準の結果が分からない状態で行われたか。　　はい/いいえ/不明 ・閾値が用いられた場合，それは事前に定義されたか。　　はい/いいえ/不明 **インデックス検査の実施または解釈はバイアスを生じた可能性があるか。**　　**リスク：低　/高　/不明** **B.適用可能性に関する懸念** **インデックス検査の実施や解釈はレビュークエスチョンと異なる懸念があるか。**　　**懸念：低　/高　/不明**

ドメイン 3：参照基準 **A. バイアスリスク**
参照基準と、それがどのように実施され解釈されたかの記述：
・参照基準は診断標的を正しく分類していると考えられるか。　　はい/いいえ/不明 ・参照基準結果の解釈は，インデックス検査結果が分からない状態で行われたか。　　はい/いいえ/不明 **参照基準の実施や解釈はバイアスを生じた可能性があるか。**　　**リスク：低　/高　/不明** **B.適用可能性に関する懸念** **参照基準により定義された診断標的はレビュークエスチョンに合致しない懸念があるか。**　　**懸念：低　/高　/不明**

ドメイン 4：フローとタイミング **A. バイアスリスク**
インデックス検査および/または参照基準を受けなかった患者，あるいは 2x2 分割表から

除外された患者の記述（フローダイアグラムを参照）： インデックス検査から参照基準までの期間およびこの期間に行われた介入の記述：
・インデックス検査と参照基準の間に適切な期間があったか。　はい/いいえ/不明 ・全ての患者が参照基準を受けたか。　はい/いいえ/不明 ・患者は同一の参照基準を受けたか。　はい/いいえ/不明 ・全ての患者が解析に含まれていたか。　はい/いいえ/不明 **患者のフローはバイアスを生じた可能性があるか。　リスク：低　/高　/不明**

【参考資料１　：QUADAS-2　記入方法】

フェーズ１：レビュークエスチョンの記述

患者（セッティング，インデックス検査の使用目的，症状，事前検査）：
インデックス検査：
参照基準および診断標的：

フェーズ２：各レビューのための調整**

フェーズ３：一次研究のフローダイアグラム作成

一次研究に掲載されたフローダイアグラムのレビューを行うが，フローダイアグラムが報告されていない，もしくは適切でない場合はレビュー著者がフローダイアグラムを作成する。

フェーズ４：バイアスリスクと適用可能性の判定

上で定義したリサーチクエスチョンに対しバイアスリスクの 4 つの主なドメインと適用可能性に関する懸念に関してそれぞれ評点付けする。バイアスリスクと適用可能性は「低」，「高」，「不明」に判定される。

ドメイン 1：患者選択

A. バイアスリスク

バイアスリスク判定の裏付けに用いる情報，シグナリングクエスチョン，バイアスリスクの判定の 3 つのセクションから構成

患者選択方法の記述：

・連続した患者かランダムサンプルを組み入れたか。　はい/いいえ/不明
・症例対照デザインではないか*。　はい/いいえ/不明
・その研究は不適切な除外を行っていないか*。　はい/いいえ/不明
患者選択はバイアスを生じた可能性があるか。　**リスク：低　/高　/不明**

B.適用可能性に関する懸念

各ドメインのバイアスおよび適用可能性の判定に役立つ一連のシグナリングクエスチョンに「はい」，「いいえ」または「不明」で回答。「はい」はバイアスリスクが低いことを示す。

組み入れられた患者の記述（レビューイング）：

組み入れられた患者はレビュークエスチョンに合致していない懸念があるか。　**懸念：低　/高　/不明**

ドメイン 2：インデックス検査

複数のインデックス検査が用いられている場合，検査毎に作成。

A. バイアスリスク

インデックス検査と，それがどのように実施され解釈されたか記述：

・インデックス検査の結果の解釈は，参照基準の結果が分からない状態で行われたか。　はい/いいえ/不明
・閾値が用いられた場合，それは事前に定義されたか。　はい/いいえ/不明
インデックス検査の実施または解釈はバイアスを生じた可能性があるか。　**リスク：低　/高　/不明**

B.適用可能性に関する懸念

インデックス検査の実施や解釈はレビュークエスチョンと異なる懸念があるか。　**懸念：低　/高　/不明**

ドメイン 3：参照基準

A. バイアスリスク

参照基準と、それがどのように実施され解釈されたかの記述：

・参照基準は診断標的を正しく分類していると考えられる　はい/いいえ/不明

か。 ・参照基準結果の解釈は，インデックス検査結果が分からない状態で行われたか。　はい/いいえ/不明 **参照基準の実施や解釈はバイアスを生じた可能性があるか。　リスク：低　/高　/不明** **B.適用可能性に関する懸念** **参照基準により定義された診断標的はレビュークエスチョンに合致しない懸念があるか。　懸念：低　/高　/不明**

ドメイン 4：フローとタイミング **A. バイアスリスク**
インデックス検査および/または参照基準を受けなかった患者，あるいは 2x2 分割表から除外された患者の記述（フローダイアグラムを参照）： インデックス検査から参照基準までの期間およびこの期間に行われた介入の記述：
・インデックス検査と参照基準の間に適切な期間があったか。　はい/いいえ/不明 ・全ての患者が参照基準を受けたか。　はい/いいえ/不明 ・患者は同一の参照基準を受けたか。　はい/いいえ/不明 ・全ての患者が解析に含まれていたか。　はい/いいえ/不明 **患者のフローはバイアスを生じた可能性があるか。　リスク：低　/高　/不明**

＊訳者注：〝症例対照デザインでない〞，〝不適切な除外を行っていない〞など，好ましい回答が〝はい〝となる。

＊＊訳者注：原文では、フェーズ 2 の「各レビューのための調整」はシートになく、以下のフェーズにずれがある。

【参考資料 2　：ベイジアン法によるメタアナリシス】

Dendukuri らによる拡張された HSROC のモデルについて述べ、それをベイジアン法によって実行するための OpenBUGS あるいは WinBUGS のコードについて解説する。

評価診断法（インデックス検査を T_1、参照基準を T_2 とし、結果が陽性の場合 1 、陰性の場合は 0 の値とする。D を疾患の有無を表す変数として、D=1 は疾患あり、D=0 は疾患なしとする。参照基準 T_2 の感度を S_2、特異度を C_2 とすると感度・特異度は次の式で表される：

T_2 の感度＝S_2=P(T_2=1 | D=1)
T_2 の特異度=C_2=P(T_2=0 | D=0)

参照基準の T_2 が完全な診断法である場合には、S_2=1.0、C_2=1.0 に設定することになる。以下に解説する下記のコードはそのように設定されている。

Z_1 を隠れ変数で連続変数とし、正規分布に従い、研究間で異なる分布とする。Z_1 は値が大きい方が疾患である確率、すなわち D=1 の確率が高くなり、小さいと疾患でない確率、すなわち D=0 の確率が高くなるとする。なお、Z_1 の取りうる値には制限があるのが普通なので、HSROC で得られる ROC 曲線はその範囲で有効と解釈する必要がある。

疾患でない場合、すなわち D=0 の場合、Z_1 は以下の正規分布に従うとする：

$$Z_1 \sim N\{-0.5\ \alpha_j, \exp(-0.5\beta)\}$$

疾患である場合、すなわち D=1 の場合、Z1 は以下の正規分布に従うとする：

$$Z_1 \sim N\{0.5\ \alpha_j, \exp(0.5\beta)\}$$

j は研究番号を表し、α_j は研究 j におけるこれら 2 つの正規分布の平均値の差に相当し、標準偏差の比が $\exp(\beta)$に相当する。α_j は診断精度の指標となり大きいほど 2 つの分布は離れていることになり、診断精度も高くなる。

α_j に対して、上位の階層を想定し、その平均値を Λ（ラムダの大文字)、分散を σ_α^2 で表すと、α_j は次の正規分布に従う：

$$\alpha_j \sim N(\Lambda, \sigma_\alpha^2)$$

この部分が階層モデルとなっており、Λ は全研究を通した平均値となり、σ_α^2 は研究間のバラツキの指標となる。階層モデルを適用すると適用しない場合に比べて、α_j が他の研究のデータの情報を取り込むことになり、ばらつきが小さくなる。これは以下の θ_j についても同様である。

それぞれの研究は異なる閾値 θ j を用いていると想定し、θ_j の上位の階層を想定して、その平均値を Θ（シータの大文字)、分散を σ_θ^2 で表すと、θ_j は次の正規分布に従う：

$$\theta_j \sim N(\Theta, \sigma_\theta^2)$$

したがって、α_j と同じく θ_j についても研究間のバラツキを想定している。しかしながら、Λ、Θ、β については研究間で共通の値を想定している。

以上から、評価診断法 Index test である T1 の感度は：

$S_{1j} = \Phi\{-(\theta_j - 0.5\alpha_j)/\exp(0.5\beta)\}$

特異度は：

$C_{1j} = \Phi\{(\theta_j + 0.5\alpha_j)/\exp(-0.5\beta)\}$

となる。なお、Φ（ファイの大文字）は図 1 の右に示すように、2 つの分布の閾値 θ_j 左右の累積確率密度を表す。

表 2. T1,T2 の結果を示す四分表

	至適基準 T_2（＋）	至適基準 T_2（－）
評価診断法 T_1（＋）	TP	FP
評価診断法 T_1 (－)	FN	TN

分かりやすくするため、研究番号を表す j の添え字を除いて表記すると（表 3)、T_1 の陽性率 t_1 = (TP+FP)/(TP+FP+FN+TN)、T_2 の陽性率 t_2= (TP+FN)/(TP+FP+FN+TN)とすると、それぞれのセルの人数を総人数に対する割合で表すと表 4 のようになる：

表 3. T1,T2 の結果を割合で表した四分表

	至適基準 T_2（＋）	至適基準 T_2（－）
評価診断法 T_1（＋）	$t_1 \cdot t_2$	$t_1 \cdot (1 - t_2)$
評価診断法 T_1 (－)	$(1 - t_1) \cdot t_2$	$(1 - t_1) \cdot (1 - t_2)$

表 4 の 4 つのセルの値の合計は $t_1 \cdot t_2 + t_1 \cdot (1 - t_2) + (1 - t_1) \cdot t_2 + (1 - t_1) \cdot (1 - t_2) = t_1 \cdot t_2 + t_1 - t_1 \cdot t_2 + t_2 - t_1 \cdot t_2 + 1 - t_2 - t_1 + t_1 \cdot t_2 = 1.0$ である。

研究 j における有病率を π_j で表し、t_1, t_2 を t_{1j}, t_{2j} で表し、参照基準 T_2 の感度 S_2、特異度 C_2 とすると、これらから、4 つのセルに入る確率 A_j、B_j、C_j、D_j は表 5 のように表される：

表 5. T2 の感度・特異度、T1 の感度・特異度と有病率から算出される 4 つのセルに入る確率。

	至適基準 T_2（＋）	至適基準 T_2（－）
評価診断法 T_1（＋）	$A_j=[\pi\Phi\{-(\theta_j-0.5\alpha_j)/\exp(0.5\beta)\}S_2 + (1-\pi)\Phi\{-(\theta_j+0.5\alpha_j)/\exp(-0.5\beta)\}(1-C_2)]\ t_{1j}{}^{\cdot}t_{2j}$	$B_j=[\pi\Phi\{-(\theta_j-0.5\alpha_j)/\exp(0.5\beta)\}(1-S_2) + (1-\pi)\Phi\{-(\theta_j+0.5\alpha_j)/\exp(-0.5\beta)\}C_2]\ t_{1j}{}^{\cdot}(1-t_{j2j})$
評価診断法 T_1（－）	$C_j=[\pi\Phi\{(\theta_j-0.5\alpha_j)/\exp(0.5\beta)\}S_2 + (1-\pi)\Phi\{(\theta_j+0.5\alpha_j)/\exp(-0.5\beta)\}(1-C_2)]\ (1-t_{1j})^{\cdot}t_{2j}$	$D_j=[\pi\Phi\{-(\theta_j-0.5\alpha_j)/\exp(0.5\beta)\}(1-S_2) + (1-\pi)\Phi\{-(\theta_j+0.5\alpha_j)/\exp(-0.5\beta)\}C_2]\ (1-t_{1j})^{\cdot}(1-t_{2j})$

なお、参照基準が完全な場合は、$S_2=1.0$, $C_2=1.0$ と設定する。

以上から、観察されたデータの尤度 Likelihood は次の式で表される：

$$L(\Theta,\Lambda,S_2,C_2,\sigma_\alpha^2,\sigma_\beta^2,\ \beta,\pi,\alpha_j,\theta_j,j=1,\ldots,J \mid t_{1j},t_{2j},j=1,\ldots,J)$$

$$=\prod_{j=1}^{J}(A_j\cdot B_j\cdot C_j\cdot D_j)$$

このモデルを、OpenBUGS あるいは WinBUGS のコードにすると下記のごとくになる。正規分布の関数 dnorm( )、正規分布の累積密度関数 phi( )、多項分布の関数 dmulti( )、ベータ分布の関数 dbeta( )、一様分布の関数 dunif( )が用いられている。変数 pi[ ]が各研究の有病率、p[1,]が各研究の感度、p[2,]が各研究の偽陽性率、prob[,]が四分表の各セルに入る確率を表している。変数 s2、c2 は参照基準（T2）の感度と特異度を表し、ここでは参照基準が完全なものであることを前提に 1 に固定されている。

このコードでは S_new、C_new で感度・特異度の予測値、theta_new、alpha_new でΘ、Λの予測値を得ている。通常のメタアナリシスの結果として用いられるのは、S_overall の中央値が感度の統合値でその 95%確信区間が 95%信頼区間、C_overall の中央値が特異度の統合値でその 95%確信区間が 95%信頼区間に相当する。

データは4つのカラムで表され、1行がひとつの研究に対応する。カラムのラベルは results[,1] results[,2] results[,3] results[,4]で表記され、TP, FP, FN, TNの人数を示す。これらは行内で1つの半角スペースで区切られ、行の最後は改行である。研究数は変数snに代入される。下記のDATAの部分を書き換えて、さまざまなデータに適用することができる。

```
model {
for(i in 1:sn) {
        theta[i] ~ dnorm(THETA,prec[1])
```

```
        alpha[i] ~ dnorm(LAMBDA,prec[2])
        p[1,i] <- phi(-(theta[i] - 0.5*alpha[i])/exp(beta/2))
        p[2,i] <- phi(-(theta[i] + 0.5*alpha[i])*exp(beta/2))
        prob[i,1] <- pi[i]*(p[1,i] * s2 ) + (1-pi[i])*( p[2,i] * (1-c2) )
        prob[i,2] <- pi[i]*(p[1,i] * (1-s2) ) + (1-pi[i])*( p[2,i] * c2 )
        prob[i,3] <- pi[i]*( (1-p[1,i]) * s2 ) + (1-pi[i])*( (1-p[2,i]) * (1-c2) )
        prob[i,4] <- pi[i]*( (1-p[1,i]) * (1-s2)) + (1-pi[i])*( (1-p[2,i]) * c2 )
        results[i,1:4] ~ dmulti(prob[i,1:4],n[i])
        n[i]<-sum(results[i,1:4])
        pi[i] ~ dbeta(1,1)
        se[i] <- p[1,i]
        sp[i] <- 1-p[2,i]
        }
for(j in 1:2) {
        prec[j] <- pow(sigma[j],-2)
        sigma[j] ~ dunif(0,2)
        }
THETA ~ dunif(-1.5,1.5)
LAMBDA ~ dunif(-3,3)
beta ~ dunif(-0.75,0.75)
S_overall<-phi(-(THETA-LAMBDA/2)/exp(beta/2))
C_overall<-phi( (THETA+LAMBDA/2)*exp(beta/2))
theta_new ~ dnorm(THETA,prec[1])
alpha_new ~ dnorm(LAMBDA,prec[2])
S_new<-phi(-(theta_new-alpha_new*0.5)/exp(beta*0.5))
C_new<-phi( (theta_new+alpha_new*0.5)*exp(beta*0.5))
s2 <- 1
c2 <- 1
}
DATA
list(sn=27)
results[,1] results[,2] results[,3] results[,4]
55 81 23 224                    #TP, FP, FN, TNのデータ、スペース区切り
48 30 11 179
51 23 8 77
・・・・・・・
END
```

Dendukuri らは参照基準が完全でない場合、研究によって参照基準が異なる場合のモデルおよび WinBUGS のコードも発表している。必要に応じてこれらを利用することも可能である。

【参考資料 3：QUADAS-2 との相違点】

表．Minds の DTA 評価法と QUADAS-2 との相違点

	Minds	QUADAS-2
ドメイン	選択バイアス、インデックス検査、参照基準、症例減少バイアス、フローとタイミング、その他。	患者の選択、インデックス検査、参照基準、フローとタイミング。
項目	各ドメインに１～２つの項目を設定。	項目の設定はなく、シグナリングクエスチョンとして設定。
Two-gate study	対象に含めない。	対象に含めてもいい？
ランダムサンプルの評価	臨床に即したランダムサンプルかどうかを対象者の属性などから総合的に判定。	連続した患者かランダムサンプルを組み入れたかで判定。
インデックス検査と参照基準の結果の判定	盲検化という用語を用いている。（内容的には同じ）。	互いの結果がわからない状態で判定が行われたかを問う。
症例減少バイアス	症例減少バイアスのドメインを設定。参照基準、インデックス検査を受けなかったり、属性のデータが不明で除外された例を評価する。	フローとタイミングのドメインの中でインデックス検査、参照基準を受けなかった患者の記述として評価する。
フローとタイミングの評価	インデックス検査と参照基準が同時期に実施されたかを評価。フローは、それぞれの患者の診療の中での流れを意味している。	インデックス検査から参照基準までの期間とその期間に行われた介入の記述も行う。フローがデータ集計時の患者数の流れを意味している。
非直接性の評価	対象、インデックス検査、参照基準、アウトカムの 4 項目について評価する。	患者の選択、インデックス検査、参照基準の 3 ドメインについて評価する。

【解説】

QUADAS-2 のドメイン 1 は患者選択で、シグナリングクエスチョンに“症例対照デザインではないか”が含まれていることから、Two-gate study を解析対象に含むことが前提とされていることがわかる。しかし、Two-gate study はその診断法が適用される患者集団を適切に代表することは難しく、特に特異度については真の値からの乖離が大きくなる可能性が高い。そのため、われわれは Single-gate study のみをメタアナリシスの対象とすることを前提としている。

また、QUADAS-2 では連続した患者であれば患者選択におけるバイアスリスクが低いと判定する構造になっているが、研究対象の患者が実際の患者集団からのランダムサンプルかどうかはこれだけでは決められず、その研究の対象者のさまざまな属性を評価して判断する必要がある。われわれは、選択バイアスとして“臨床に即したランダムサンプル”かどうかを評価するようにしている。

QUADAS-2 ではドメイン 2 のインデックス検査とドメイン 3 の参照基準において、シグナリングクエスチョンとして“それぞれの結果をわからない状態で結果の解釈が行われたか”どうかを評価するようになっている。これに対して、われわれは盲検化という表現を用いている。

また QUADAS-2 のドメイン 4 のフローとタイミングでは、“全ての患者が解析に含まれていたか。”のシグナリングクエスチョンが含まれているが、これは症例減少バイアスの概念に相当するので、われわれは症例減少バイアスとして評価するようにしている。フローとタイミングとしてはインデックス検査と参照基準が同時期に実施されたかどうかを評するようにしている。同時期とは間に治療的介入が行われたり、急性の疾患で時期が変わったりあるいは悪性腫瘍のように時間経過で増大して、陽性率が変化することがないかを評価することである。

QUADAS-2 では適用可能性をドメイン 1 から 3 について評価するようになっており、ドメイン 4 については適用可能性の項はない。われわれは、非直接性として対象、インデックス検査、参照基準、アウトカムの 4 項目を評価するようにしている。前 3 者については、QUADAS-2 と同様であるが、アウトカムは QUADAS-2 とは異なる設定である。CQ のアウトカムの設定についてはすでに述べたとおりであるが、たとえば、診断標的をアウトカムとして設定する場合、病理学的診断を根拠とするか、臨床的診断と画像診断を根拠とするかで非直接性が異なる。これは“不完全な参照基準”とは異なる意味をもつので、非直接性として評価する必要がある。

アウトカムの非直接性については、GRADE の診断の CPG 作成方法では評価項目に設定されていない。彼らの方法では、アウトカムとして、TP, FN, TN, FP、あいまいな結果、検査に伴う害を一律に設定することになっており、非直接性の評価は、対象、インデックス検査、参照基準についてのみ行うことになっている。

参考文献

1) Tatsioni A, Zarin DA, Aronson N, Samson DJ, Flamm CR, Schmid C, Lau J: Challenges in systematic reviews of diagnostic technologies. Ann Intern Med 2005;142:1048-55.
2) Jonathan Hsu, Jan L Broz3 ek, Luigi Terracciano, Julia Kreis, Enrico Compalati, Airton Tetelbom Stein, Alessandro Fiocchi, Holger J Schünemann. Application of GRADE : Making evidence-based recommendations about diagnostic tests in clinical practice guidelines. Implementation Science, 2011 ; 6:62.
3) Schünemann HJ, Oxman AD, Brozek J, Glasziou P, Jaeschke R, Vist GE, Williams JW Jr, Kunz R, Craig J, Montori VM, Bossuyt P, Guyatt GH; GRADE Working Group. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. BMJ. 2008; 336(7653): 1106-10.
4) Schünemann HJ, Oxman AD, Brozek J, Glasziou P, Bossuyt P, Chang S, Muti P, Jaeschke R, Guyatt GH. GRADE: assessing the quality of evidence for diagnostic recommendations. Evid Based Med. 2008; 13(6): 162-3.
5) Berkman ND, Lohr KN, Morgan LC, Richmond E, Kuo TM, Morton S, Viswanathan M, Kamerow D, West S, Tant E. Reliability Testing of the AHRQ EPC Approach to Grading the Strength of Evidence in Comparative Effectiveness Reviews. Methods Research Report. (Prepared by RTI International–University of North Carolina Evidence-based Practice Center under Contract No. 290-2007-10056-I.) AHRQ Publication No. 12-EHC067-EF. Rockville, MD: Agency for Healthcare Research and Quality. May 2012. www.effectivehealthcare.ahrq.gov/reports/final.cfm.
6) Leeflang MM, Deeks JJ, Gatsonis C, Bossuyt PM. Systematic reviews of diagnostic test accuracy. Ann Intern Med. 2008; 149(12): 889-97.
7) Thornbury JR: Eugene W. Caldwell Lecture. Clinical efficacy of diagnostic imaging: love it or leave it. AJR 1994;162:1-8.PMID:8273645
8) Haynes RB, Wilczynski NL: Optimal search strategies for retrieving scientifically strong studies of diagnosis from Medline: analytical survey. BMJ. 2004 1;328:1040.PMID: 15073027
9) Beynon R, Leeflang MM, McDonald S, Eisinga A, Mitchell RL, Whiting P, Glanville JM: Search strategies to identify diagnostic accuracy studies in MEDLINE and EMBASE. Cochrane Database Syst Rev. 2013;9:MR000022. pub3.
10) Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, Moher D, Rennie D, de Vet HC, Lijmer JG; Standards for Reporting of Diagnostic Accuracy. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Ann Intern Med. 2003 Jan 7;138(1):W1-12.

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 18;155(8):529-36.
12) http://srdta.cochrane.org/handbook-dta-reviews
13) Sox HC, Higgins MC, Owens DK: Medical decision making. 2013, Wiley-Blackwell, West Sussex, UK.
14) Moses LE, Shapiro D, Littenberg B: Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. Stat Med 1993;12:1293-316.
15) Littenberg B, Moses LE: Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytic method. Med Decis Making 1993;13:313-21.
16) Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH: Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. J Clin Epidemiol 2005;58:982-90.
17) http://cran.r-project.org/web/packages/mada/index.html
18) Rutter CM, Gatsonis CA: A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Stat Med 2001;20:2865-84.
19) Deeks JJ, Macaskill P, Irwig L: The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. J Clin Epidemiol 2005; 58:882 - 893.
20) 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;68:1285-93.
21) http://www.nandinidendukuri.com/
22) Dahabreh IJ, Trikalinos TA, Lau J, Schmid C. An Empirical Assessment of Bivariate Methods for Meta-Analysis of Test Accuracy. Methods Research Report. (Prepared by Tufts Evidence-based Practice Center under Contract No. 290-2007-10055-I.) AHRQ Publication No 12(13)-EHC136-EF. Rockville, MD: Agency for Healthcare Research and Quality. November 2012. www.effectivehealthcare.ahrq.gov/reports/final/cfm.

引用記載例：

森實敏夫，河合富士美，小島原典子．“特別寄稿 5，診断に関する診療ガイドライン（CPG）の作成”．Minds 診療ガイドライン作成マニュアル．小島原典子，中山健夫，森實敏夫，山口直人，吉田雅博編．公益財団法人日本医療機能評価機構．2015，http://minds4.jcqhc.or.jp/minds/guideline/special_articles5.pdf, 2015 年 12 月 15 日参照.