# 診断精度研究の系統的レビューとメタアナリシス

杣取恵太[1]・坂本次郎[1]・時田椋子[1]・鈴木彩夏[1]・国里愛彦[2]

# Systematic Reviews and Meta-analysis of Diagnostic Test Accuracy

Keita Somatori[1], Jiro Sakamoto[1], Ryoko Tokita[1], Ayaka Suzuki[1] and Yoshihiko Kunisato[2]

**Abstract**：Diagnostic tests specify whether a person has a specific disease or not and it extremely contribute decision-making of the intervention. Various types of diagnostic test are proposed and their methodological qualities have been improved. However, there are much inconsistent evidences in diagnostic test accuracy studies. We have become so difficulty of decision making in diagnostic test. Therefore, we believe that individual studies on diagnostic test accuracy should be synthesized for evidence based clinical psychology. In systematic reviews of diagnostic test accuracy, Cochrane collaboration is making up its guideline “Handbook for diagnostic test accuracy reviews”. In the handbook, some key components of systematic reviews and meta-analysis in diagnostic test accuracy are explained, which contain the drawing up protocol, search strategy, assessing methodological quality, and meta-analysis. We review the key components of systematic reviews and meta-analysis according to “Handbook for diagnostic test accuracy reviews”.

**Keywords**：diagnostic test accuracy, meta-analysis, systematic reviews, Cochrane review, QUADAS-2

## 1. はじめに

医療実践において，エビデンスに基づく医療（Evidence Based Medicine）という言葉が最初に用いられ始めてから四半世紀近くが経過した。EBM とは，個々の患者のケアにおける意思決定のために，現状において最善のエビデンスを慎重かつ明示的に思慮深く用いることである（Sackett, Rosenberg, Gray, Haynes & Richardson, 1996）。身体疾患を対象とした医療実践だけでなく，臨床心理実践においても，同様にエビデンス（根拠）に基づいた臨床心理実践（ Evidence Based Clinical Psychology: EBCP）を行う必要性が指摘されてきている（丹野，2001）。患者のケアに関する意思決定においては，まず患者の診断・アセスメントに関するエビデンス，患者の予後予測や発症にかかわる要因に関するエビデンス，その疾患に有効な治療法に関するエビデンスなどがある。本稿で扱う診断検査（diagnostic test）は，患者の診断・アセスメントのエビデンスに関わり，患者の状態を特定し，介入計画を立てる上で必要不可欠なものである。特定の疾患の予後・発症要因・治療法に関するエビデンスは，正確な診断やアセスメントの上に成り立っており，EBCP において，診断検査の診断精度を調べた研究は重要になる。

診断検査には，質問紙，投影法検査，神経心理学的検査や脳画像検査など患者の健康状態について何らかの情報を与えるすべての検査が含まれる。このような診断検査は非常に多岐にわたり，新たな検査も次々と生み出されつつある。そのため，われわれは目の前の患者に対してどの検査を用いるべきか判断に悩まされることもしばしば生じる。White, Schultz, & Enuameh（2011）は診断検査に求められる主要な要素として以下の三つを挙げている。一つは検査の速さである。迅速に実施でき，結果を素早く導出できることによって患者に対して早い段階で治療が可能になる。二つ目は，検査実施にかかわるコストの低さである。低コストであることによって多くの患者が利用でき，幅広い対象者が治療を受けることができる。三つ目は簡便さである。簡便な検査であることによって検査を行う上でのエラーや解釈におけるエラーが減少し，患者にとっても信頼できる検査になりうる。

既存の検査よりも患者にとって有益な検査を作成し，その検査の診断精度（Diagnostic Test Accuracy:

受稿日2014年10月９日　受理日2014年11月11日

1　専修大学大学院文学研究科（Graduate School of the Humanities, Senshu University）
2　専修大学人間科学部心理学科（Department of Psychology, Senshu University）

DTA）を検討する研究が近年増加してきている（White et al., 2011）。White et al.（2011）は完璧な検査など存在せず，故に常に現行の検査を上回る検査を生み出していく必要があるとしている。このためにも，既存の検査と新規の検査の精度を比較する研究は不可欠であり，また診断精度研究によって現場においてどの検査を選択すべきかに示唆を与えることができるといえる。

一方，診断精度研究は増えてきているものの，検査対象者の属性やサンプルサイズの違い，検査者の解釈の違いなどによる研究間の結果のばらつきの問題もある。そこで，それらの研究知見を系統的にレビューしたり，結果を統合するメタアナリシスを実施する必要がある。診断精度研究の系統的レビューやメタアナリシスを行うことにより，個々の研究のエビデンスを統合した診断精度を推定することができる。また，異質性の評価を通して，個々の研究の結果をばらつかせている要因についても検討することができる。

しかしながら，診断精度に関するメタアナリシス研究は介入研究に関するメタアナリシス研究などに比較して研究が遅れていた。その原因としては，(1)各研究の結果を統合する統計手法や研究間のばらつきを評価する手法の開発が不十分という方法論上の問題，(2)診断精度のメタアナリシスを報告する際どのような情報を載せるべきかという報告の方法に関するガイドラインが定まっていない点がある。以上をうけて，White et al.（2011）やBuntinx, Aertgeerts, & Macaskill（2009）などのような診断精度研究の系統的レビューやメタアナリシスに関するガイドラインを記述した書籍も発刊されるようになってきている。また，健康に関するさまざまな臨床研究の収集・分析・統合に関する知見をまとめているコクラン共同計画（Cochrane collaboration）の Diagnostic Test Accuracy Working Group（http://srdta.cochrane.org/）が「Handbook for DTA reviews」という診断精度のメタアナリシスに関するガイドラインを現在作成している。「Handbook for DTA reviews」は，診断精度研究のメタアナリシスにおいて今後重要な枠組みになることが想定される。そこで，本稿では主に現在（2014年10月）において作成中の「Handbook for DTA reviews」のガイドラインを中心に診断精度研究の系統的レビューとメタアナリシスの手法について解説する。

## 2． 診断精度研究について

### 2.1 指標検査（Index test）と参照基準（Reference standard）

本章では主に「Handbook for DTA reviews」および White et al.（2011）に従って診断精度の一次研究において報告されるべき事項について解説する。診断精度とは，その検査が患者の状態を正確に反映できる度合いとして定義される。研究において関心のある診断検査は指標検査（Index test）とよばれる。指標検査の結果は主に陽性・陰性の２値，あるいは「非常に悪い」～「非常に良い」などの段階的なもの，０点～100点などの幅で得点を取りうる連続的なものの３種類のいずれかの形であらわされる。なお，段階的・連続的な結果もカットオフ得点（閾値）を用いて最終的には陽性・陰性の２値に振り分けられる。

指標検査の結果に対して患者の真の状態を表すのが参照基準（Reference standard）である。実際にはその患者が真に疾患を有しているか否かを完全に特定することはできない。そこで，現存する最も精度が高い診断検査の結果をその患者の真の状態と「仮定」した上で指標検査の精度を求める。すなわち，指標検査に基づく陽性・陰性があり，「仮定」された真の状態として参照基準に基づく陽性・陰性がある。そこで，これらは Table 1 のような２×２のマトリックスで表現される。なお，真の状態として「参照基準」ではなく「至適基準（Gold standard）」という用語が用いられることもある。しかしながら，これは患者の真の状態を測定しているという誤解を生む恐れがあるため，適切な表現ではないとされている（Virgili, Conti, Murro, Gensini, & Gusinu, 2009）。本稿においても，Virgili et al.（2009）に従って「参照基準」を用いる。

### 2.2 感度（Sensitivity）と特異度（Specificity）

診断検査の精度は，指標検査および参照基準を用いることで患者を Table 1 に示すような２×２のマトリックスのいずれかのセルに振り分け，感度・特異度と呼ばれる二つの指標によって検討される。感度とは，参照基準が陽性であった場合に指標検査の結果が陽性である割合である。感度は，以下の式で算出され，患者を正しく陽性だと診断できる程度を表し，高い方が望ましいとされる。

$$感度 = \frac{TP}{TP + FN} \quad (1)$$

特異度とは，参照基準が陰性であった場合に指標検査の結果が陰性である割合のことである。特異度は，以下の式で算出され，患者でない者を正しく陰性だと診断できる程度を表し，こちらも高い方が望ましいとされる。

$$特異度 = \frac{TN}{FP + TN} \quad (2)$$

感度・特異度は診断精度の最も基本的な指標であり，上記の各セルの値とともに必ず報告されなければならない。このように，診断精度研究においては二つの指標を用いて精度を記述する。感度・特異度以外の指標としては，陽性的中率（Positive Predictive Value: PPV）と陰性的中率（Negative Predictive Value: NPV）もあり，以下の式で算出される。

$$PPV = \frac{TP}{TP + FP} \quad (3)$$

$$\mathrm{NPV} = \frac{TN}{TN + FN} \quad (4)$$

PPV と NPV は，それぞれ指標検査が陽性であった場合の真の陽性者の割合，指標検査が陰性であった場合の真の陰性者の割合と定義される。なお，以降で述べる尤度比や ROC 曲線は感度・特異度を元に算出されるため，感度・特異度を報告する方がより一般的であると考えられる。

### 2.3　陽性尤度比（Positive likelihood ratios: LR（+））と陰性尤度比（Negative likelihood ratios: LR（−））

陽性尤度比とは真の状態が陰性である人よりも真の状態が陽性である人の方が何倍指標検査で陽性になりやすいかを表している。以下の式で算出され，陽性尤度比が高いほど，検査の結果が陽性であれば真の状態も陽性である可能性が高く，確定診断につながるとされる（杉岡・野口・大西，2014）。

$$LR（+） = \frac{感度}{1 - 特異度} \quad (5)$$

一方，陰性尤度比とは真の状態が陰性である人よりも真の状態が陽性である人の方が何倍指標検査で陰性になりやすいかを表している。以下の式で算出され，陰性尤度比が低いほど，検査の結果が陰性であれば真の状態が陰性である可能性が高く，除外診断につながるとされる（杉岡ほか，2014）。

$$LR（−） = \frac{1 - 感度}{特異度} \quad (6)$$

尤度比は，臨床家にとって感度や特異度よりも解釈しやすい便利な指標とされる。臨床実践において，臨床家は，検査をする前に，ある程度目の前の患者がその疾患である確率を想定している（検査前確率）。以下の式を用いて，検査前確率をオッズにした検査前オッズを算出し，その検査前オッズに尤度比を掛けると検査後オッズを計算することができる（詳しくは，古川，2000を参照）。尤度比によって臨床実践に則した検査の利用が可能になる。

$$検査後オッズ = 尤度比 \times 検査前オッズ \quad (7)$$

$$オッズ = \frac{その疾患を有している確率}{その疾患を有していない確率} \quad (8)$$

なお，陽性尤度比と陰性尤度比をまとめる形で一つの指標を用いて表すことも可能である。診断オッズ比（Diagnostic Odds Ratio: DOR）は，以下の式で算出され，高い方が望ましいとされる。

$$DOR = LR（+）／LR（−） \quad (9)$$

DOR は，これまでとは異なり，一つの指標で診断精度について議論できるため，シンプルかつ統計的に扱いやすい指標であると言える。一方で，同じ診断オッズ比でも異なる感度・特異度の組み合わせを取りうるため，診断オッズ比だけでは臨床適用についてあまり言及できないという問題もある。

### 2.4　閾値（thresholds）

閾値は段階的・連続的な値を取りうる検査の結果を陽性・陰性の 2 値に振り分ける基準である。同じ検査においても，閾値が異なれば感度・特異度が異なる。そのため，指標検査，参照基準のどちらにおいても閾値は必ず報告しなければならない。

閾値は感度・特異度と非常に密接な関係にある。Fig-

Table 1　指標検査の結果および真の状態

| | | 真の状態（Reference standard） | |
|---|---|---|---|
| | | 疾患有り | 疾患無し |
| 指標検査（Index test）の結果 | 陽性 | 真陽性(True Positive: TP) | 偽陽性(False Positive: FP) |
| | 陰性 | 偽陰性(False Negative: FN) | 真陰性(True Negative: TN) |

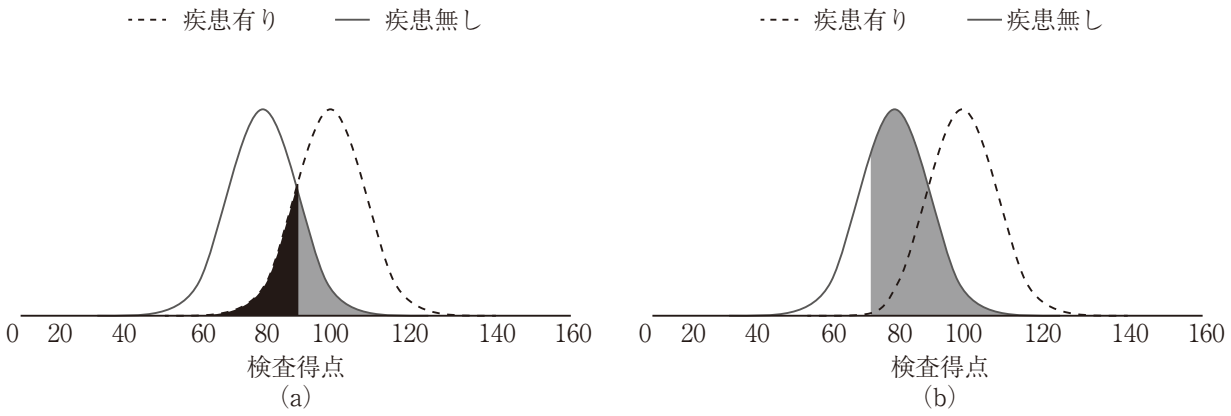


Figure 1　閾値と感度・特異度の関連

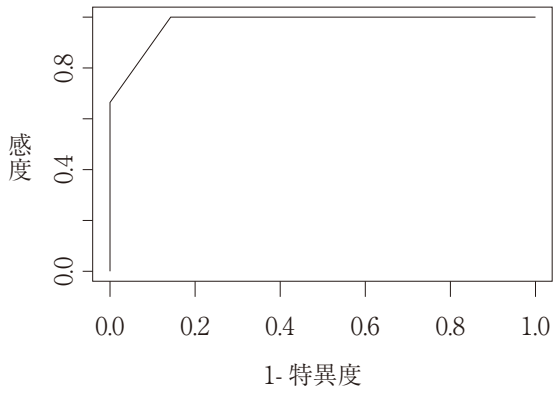


Figure 2　ROC 曲線（例）

ure 1 の(a)と(b)は，参照基準の結果が陽性である疾患有りの集団と参照基準の結果が陰性である疾患無しの集団に対して同じ分布を仮定し，異なる閾値を設定したものである。(a)の場合は感度・特異度が同程度であるが，(b)の場合は感度が非常に高く，特異度が低くなっている。閾値が変わることによって，２×２のマトリックスの各セルの値が変動し，その結果として感度・特異度も変化する。

### 2.5　ROC 曲線（Receiver Operator Characteristic Curve）

一次研究において，複数の閾値を用いて検査の感度・特異度を検討する場合に ROC 曲線が用いられる。ROC 曲線とは縦軸を感度，横軸を１－特異度として閾値が取りうるすべての値において感度・特異度を求めてプロットしたものである（Figure 2）。診断オッズ比同様，診断検査の精度を一つの指標を用いて表すものであり，閾値の値によって感度・特異度がどのように変化するかを検討することができる。

ROC 曲線の形や感度・特異度は疾患有りの集団と疾患無しの集団における指標検査得点の分布の重なる程度に大きく依存する。すなわち，疾患有りの集団と疾患無しの集団の分布があまり重ならない場合，ROC 曲線では感度・特異度ともに高い値を取りうる。しかしながら，疾患有りの集団と疾患無しの集団の分布が大きく重なってしまう場合，感度・特異度ともに低い値となってしまう。このように，疾患有りと疾患無しを識別できる程度が ROC 曲線下の面積（Area Under the Curve: AUC）に反映されるため，ROC 曲線を用いて診断検査の総合的な評価を行うことが可能である。

## 3.　診断精度研究の系統的レビューとメタアナリシス

White et al.（2011）は診断精度の系統的レビューとメタアナリシスの目的として，診断検査における一次研究のエビデンスの統合および，一次研究の質の評価を挙げている。一次研究の質の評価とは研究間のばらつきの評価およびばらつきに影響を与えている要因の検討や，一次研究の報告の質についても言及することである。系統的レビューが行われることにより，臨床適用のみならず今後の一次研究の質の向上にも貢献するとされている。

本章ではまず，(1)診断精度のメタアナリシスを行う際のプロトコル作成，(2)一次研究を収集する際の適格・除外基準，(3)一次研究を収集する際に活用するデータベースやその検索方法，(4)収集された一次研究の質の評価，(5)統計的なエビデンスの統合方法，(6)分析によって得られた結果の解釈について解説する。

### 3.1　研究の目的やプロトコルの作成

現在作成中の「Handbook for DTA reviews」では診断精度のメタアナリシスを行うにあたって事前に研究全体のプロトコルを作成し，公開することを推奨している

Table 2　診断精度のメタ分析研究におけるプロトコルのフォーマット

| | |
|---|---|
| 題目（P） | |
| 著者の詳細（P） | |
| 連絡担当者（P） | |
| 日付 | 最終改定日（P）；文献を検索した日（P）；研究が次のステップに移行する予定日（P）；プロトコルの初公開日（P）；レビューの初公開日；最新の引用文献 |
| 新着情報 | |
| ここまでの経緯 | |
| アブストラクト | 研究の背景；目的；文献の検索方法；文献の適格基準；データ収集と分析；結果；考察 |
| 要約 | ターゲット症状（P）；指標検査（P）；治療プロセス（参照基準（P）；指標検査の役割（P）；ほかの検査（P））；理論的な根拠 |
| 目的 | 副次的な目的（P） |
| 方法（P） | 一次研究を収集する際の基準（研究の種類（P）；参加者（P）；指標検査（P）；ターゲット症状（P）；参照基準（P））<br>一次研究の検索方法（電子データベース（P）；その他（P））<br>データの統合方法（P）；異質性の評価方法（P）；感度分析（P）；バイアスの評価方法（P）） |
| 結果 | 文献検索の結果；方法論の評価の結果；文献収集を行って明らかになったこと |
| 考察 | 主な結果の要約；レビューの長所・短所；レビューで得られた知見の適用可能性 |
| 著者の考察 | |
| 謝辞（P） | 臨床への示唆；研究への示唆 |
| 著者の貢献（P） | |
| 関心のある知見（P） | |
| プロトコルのレビューの内容的な差異 | |
| 付記すべきこと（published notes） | |
| サポート資源 | 内的資源；外的資源 |
| フィードバック | |
| 付録 | 検索方略（P）；QUADAS/QUADAS-2の評価結果（P） |
| 表 | 研究の特徴（適格研究の特徴；除外研究の特徴；分類不可だった研究の特徴；進行中の研究の特徴） |
| 得られた知見の表 | 適格研究；除外研究；分類不可だった研究；進行中の研究 |
| 付加的な表 | 背景；レビューの別バージョン；上記すべてに当てはまらない内容 |
| 参考にした研究 | |
| その他参照すべきこと | |
| データおよび分析結果 | |
| 図 | |

※ P は必須項目
※ Deeks et al.（2013）を元に作成

Table 3 問題の定式化

| PECO | |
|---|---|
| Population（Participants） | 対象となる参加者の属性（性別や年齢，人種） |
| Exposure（Intervention） | 診断精度を検討したい検査 |
| Comparison | 比較対象となる既存検査 |
| Outcome | 診断精度の指標 |

| PIRATE | |
|---|---|
| Population（Participants） | 対象となる参加者の属性（性別や年齢，人種） |
| Index test | 診断精度を検討したい検査 |
| Reference test | 比較対象となる既存検査 |
| Accuracy Methods | 診断精度の指標 |
| Test cut off point | 陽性・陰性となる基準 |
| Expected test use | 指標検査を用いることによって期待される効果（検査の簡便化，低コスト化など） |

（Deeks, Wisniewski, & Davenport, 2013）。研究のプロトコルは研究の目的を明確にし，文献の収集方法や結果の統合の仕方などについて事前に定義しておくものであり，系統的なレビューには必要不可欠なものである。プロトコルの内容が実際にデータの収集に当たる前に整理され，公開されることによってレビューの透明性や研究の再現性を保障することができる。また，レビューの読者に対し，その発見に至るプロセスについて理解を助ける点でも有用である（White et al., 2011）。

報告するプロトコルの内容については，コクラン共同計画によってそのフォーマット（Table 2）が作成されている（Deeks et al., 2013）。研究によって報告されるプロトコル内容が研究ごとに異なることは望ましくないため，診断精度のメタ分析を行う際にはコクラン共同計画のフォーマットに従うことが望ましい。本稿では，コクラン共同計画が作成したフォーマットに従い，研究の実施前に予め定義しておくべき事柄について解説する。

コクラン共同計画が作成したフォーマットでは日付も重要な項目の一つとされている。収集された文献はいつの段階でのことなのか，レビューはいつ公開されるのか，現在アップロードされている情報はいつのものなのか，レビューの読み手にとって非常に重要な情報である。また，レビューのプロセスの透明化を図る上でも重要なため，厳正に公開されなければならない。なお，コクラン共同計画では，系統的レビューにおけるプロトコルの作成から文献管理，データ解析，そしてレビューの維持を行うソフトの Review Manager（RevMan, http://tech.cochrane.org/revman）を公開している。RevMan でプロトコルを作成すればプロトコルを更新した際，日付が自動的に入力されるようになっている。

系統的レビューでは，系統的に文献を収集した上で分析にあたる必要がある。そのためには研究の背景，目的，方法について，あらかじめ明確にしておき，プロトコルに記載することが求められる。リサーチクエスチョンを明確にし，一定の枠組みを作成した上で文献の収集，分析，執筆にあたることでレビューの質を向上させ，読み手にとっても明瞭な理解を助けることができる。

レビューの背景，および目的について明確にするためには問題の定式化を行う必要がある。問題の定式化とは研究の主となる要素について整理することであり，PECO という枠組みを用いるのが一般的である（Table 3 左）。杉岡ほか（2014）では PECO を用いた診断精度研究における問題の定式化を推奨している。しかし，White et al.（2011）ではさらに細分化した PIRATE も推奨している（Table 3 右）。いずれの枠組みにおいても，対象となる参加者や問題となる診断検査，診断精度指標について明確な定義を行う必要がある。問題の定式化を行うことは，レビューの背景や目的を明確にする上で役立つだけではなく，後述する文献の適格基準や除外基準を作成する上でも有用である。

文献の収集方法，およびデータの統合方法についても，明確に定義しておく必要がある。文献をどのように収集したのか，適格・除外基準は何かなど，そのプロセスや基準について明記しなければならない。また，収集したデータの統合方法，結果のばらつきやバイアスの評価方法についてもあらかじめプランを立てておく必要がある。また，これらの内容については変更があった際に随時プロトコルを更新し，どの段階で，どのような理由から変更を行ったのかを記述する必要がある。このように，簡潔かつ網羅的な情報が記載されており，綿密な更新が行われることによって，レビューの系統性や厳密性

が担保されるのである。なお，オーサーシップについても，プロトコルの段階で決めておくことが望ましい。

### 3.2 研究の適格・除外基準の設定

診断精度研究における感度・特異度は，その検査を行う状況や対象者の属性などによって変動しうるものである。そのため，関心のある指標検査を用いた一次研究を無秩序に収集してしまえば，正確な診断精度を推定することはできない。そこで，主となる目的ないしは副次的な目的に従った適格・除外基準を設定し，検索された文献を振るいにかける必要がある。そこで，本節では文献収集の際の適格基準あるいは除外基準作成にあたって，Bossuyt & Leeflang（2008）を参考に留意すべき事項について述べる。

適格基準・除外基準作成にあたっては研究の種類，対象者，指標検査，参照基準，ターゲット症状について明確にしておく必要がある。診断精度研究は，検査による予後予測研究とは別のものであり，その研究デザインは基本的に横断研究になる。診断精度研究では，すでに陽性の診断を受けている患者と健常者に対し指標検査を実施して精度を検討するという研究デザインがあり，ケース・コントロール型診断精度研究と呼ばれる。これは，厳密な意味でのケース・コントロール研究というわけではなく，横断研究であるがケース・コントロール的なデザインという意味になる。なお，ケース・コントロール型診断精度研究では，適格基準の適用が 2 段階なので，Two-gate 研究とも呼ばれる。ケース・コントロール型研究の場合，参加者を集めやすい利点があるが，一般に過剰に高い感度・特異度が得られてしまうためバイアスが混入しやすいと言われている（杉岡ほか，2014）。

一方で，現時点では対象疾患を有しているかどうか不明な集団を対象に検査を行い，疾患を有するか否かを検討する研究デザインもあり，コホート型精度研究と呼ばれる。臨床現場においては，多くの場合特定の疾患を有しているかどうか疑わしい人に対し検査を行うため，検査には陽性・陰性を判定する精度が求められる（杉岡ほか，2014）。そこで，対象者の「今」の状態が疾患を有していると断定できない状態における診断精度を検討することが望ましいとされている。どちらの手法にも利点と欠点が存在するため，文献収集を行う際はどのような状況を想定した診断検査の精度を検討したいのかを明確にしておく必要がある。なお，ケース・コントロール型やコホート型という言い回しは，縦断研究の印象を抱かせるため，Two-gate や Single-gate という言い回しの方が混乱は少ないかもしれない。

同様に，対象者についても明確な基準を用いて一次研究の収集を行う必要がある。対象者に関する基準とは，例えば性別や年齢，人種や健康状態などのことを指す。すなわち，どのような母集団を想定した診断精度を検討するかである。White et al.（2011）では母集団を想定することは，臨床実践において非常に重要であるとしている。異なる母集団を用いて行った推定は不正確であり，現場の誤った解釈を引き起こす可能性があるためである。

そこで，ターゲット症状の定義が性別や人種によって異なる場合や診断のプロセスが年齢で異なる場合，健康状態によって異なる検査得点分布が仮定できる場合など，対象者についても適格基準を設定する必要がある。ターゲットとなる疾患は臨床的，方法論的，あるいは広いものから狭いものまでさまざまな形で定義される（Bossuyt & Leeflang, 2008）。レビューにおいては具体的に一つの疾患を想定し，その程度についても明確な定義が必要である。これは，臨床適用において検査の有効な範囲を限定するものであり，過度に一般化しすぎないようにするためである。

ターゲット症状によっては確立された参照基準が存在せず，複数の検査結果を総合した結果などを参照基準として採用している場合もある（杉岡ほか，2014）。ところが，参照基準がばらついていると，当然ながら感度・特異度が変動してしまう。そのため，収集する一次研究の参照基準は基本的に一つに絞られるべきであるとされている（Bossuyt & Leeflang, 2008）。一方で，指標検査においても研究間で閾値が異なり，感度・特異度がばらつく可能性がある。もちろん，閾値以外にも陽性・陰性を判断する基準が研究間で異なるケースもある。しかしながら，閾値効果の場合はデータを統合する際にある程度判別可能なため，過度に厳しい制限を設けるべきでないとされている（Bussuyt & Leeflang, 2008）。

ここまで述べてきた以外にも，場合によっては，診断精度の指標や診断精度研究の目的などを適格基準に含める必要がある（White et al., 2011）。適格基準の設定は，問題の定式化によって明確化し，レビューの目的に適した文献を収集する上で重要である。想定されるレビューの適用可能範囲を限定し，有益な一次研究のみを収集するためには，あらかじめ基準を設定してから文献検索にあたる必要がある。また，本節で述べた適格基準は，先に述べた研究プロトコルの必須事項にも含まれている。よって，文献収集から採否に至るまでのプロセス

は，第三者によって再現可能なものでなければならない。Buntinx et al.（2009）は，収集された文献を適格基準あるいは除外基準に従って振り分ける作業は，２名の評価者によってタイトルおよびアブストラクトを精査べきであるとしている。採否について両者の合意が得られない場合は，第三者による決定を仰ぐか，全文を検討することによって決定する。このことからもわかる通り，研究の採否は厳密かつ再現性の高い方法で決定されなければならない。

### 3.3 文献の検索

系統的レビューに含める文献の検索は，系統的レビューにおいて最も重要な手続きになる。本節では，de Vet, Eisinga, Riphagen, Aertgeerts, & Pewsner（2008）を参考に，文献検索の手法および参照データベースを解説する。まず，文献検索では，包括的な文献検索を行い，出版バイアスを代表とする潜在的なバイアスを可能な限り排除することが求められる。しかし，診断精度研究における出版バイアスの評価方法は確立しておらず（Song, Khan, Dinners, & Sutton, 2002），バイアス軽減のために複数のデータベースを利用することが推奨されている（Whiting, Westwood, Burke, Sterne, & Glanville, 2008）。

ここでは，MEDLINE と EMBASE の二つのデータベースを紹介する。MEDLINE は米国の医学系学術データベースであり，1950年以降に発刊されたおよそ5000種の雑誌，約160万件の文献が検索可能となっている。MEDLINE は無料版データベースである PubMed を公開しており，MEDLINE では検索されない文献が一部登録されている。EMBASE は生理医学系のデータベースであり，1974年以降に発刊されたおよそ4800種の雑誌に掲載された約120万件の文献が登録されている。

MEDLINE および EMBASE の両データベースを利用した場合には包括性の高い文献検索が可能となる（Fraser, Mowatt, Siddiqui, & Burr, 2006）。その一方で，片方のデータベースのみを用いた場合には収集される文献に偏りが生じることは不可避となる（Smith, Darzims, Quimn, & Heller, 1992; Fraser et al., 2006; Whiting et

Table 4　MEDLINE, EMBASE を除く目的別データベース（deVet et al.（2008）を元に作成）

| 目的 | 対象 | データベース名 |
|---|---|---|
| 地域別検索 | アフリカ | African Index Medicus |
| | オーストラリア | Australasian Medical Index（fee-based） |
| | 中華人民共和国 | Chinese Biomedical Literature Database（CBM） |
| | 東地中海地域 | Index Medicusfor Eastern Mediterranean Region |
| | ヨーロッパ | PASCAL（fee-based） |
| | インド | IndMED |
| | ウクライナおよびロシア | LILACS 0 |
| | 韓国 | KoreaMed |
| | 中南米 | Index Medicusfor Sout h -East Asia Region（IMSAER） |
| | 東南アジア | Panteleimon |
| | 大西洋沿岸 | Western Pacific Region Index Medicus（WPRIM） |
| 特定領域の検索 | グローバルヘルス | Global Health |
| | 看護関連 | Allied and Complementary Medicine（AMED）; British Nursing Index（BNI）; Cumulative Index to Nursing and Allied Health（CINAHL） |
| | プライマリーケア | Essential EvibencePlus |
| | 社会科学、心理学、精神医学 | Applied Social Science Index and Abstracts（ASSIA）;PsycINFO;Sociological Abstracts |
| | 全般的検索 | Google Scholar;Intute;Turning Research intoPractice（TRIP）database |
| その他の検索 | 学位論文 | ProQuest Dissertaations& Theses Database;Index to Theses in Great Britain and Ireland;DissOnline |
| | 灰色論文 | the European Association for Grey Literature Exploitation（EAGLE）;OpenSIGLE;NTIS |

al., 2008)。従って，de Vet et al.（2008）では，診断精度研究の系統的レビューのためにMEDLINEとEMBASEの二つのデータベースを利用することが推奨されている。また，特定の領域（生理・化学，看護など）に限った診断検査を対象としてレビューを行う場合や，灰色文献，学位論文なども含めた文献収集を行う際には，それらの検索に特化したデータベースを補足的に利用することが有効とされている。これら目的ごとのデータベースについてはTable 4にまとめている。また，電子化されていない文献を収集するハンドサーチや，抽出された文献の参考文献リスト参照も有効な手段となる。

具体的に文献検索を進めるにあたって，関連論文間での一貫性のない用語使用といった検索上の問題に直面することがある。このことは，必要な関連論文を的確に抽出し，取りこぼすリスクの最小化を目指す文献検索の障壁となる。従って，文献検索の洗練化のため関連論文を適切に拾い上げるための検索ワードの設定が重要となる。de Vet et al.（2008）では，検索ワード設定の際に指標検査とターゲット症状の二つを主に使用することが望ましいとされている。

検索ワードの選定には，指標検査およびターゲット症状の類義語や同義語をあらかじめ無制限に抽出しておくことが望まれる。

MEDLINEでは，データベース内で文献に付与される検索キーワードとしてMeSH（Medical Subject Headings）タームを定めており，MeSHタームの利用は指標検査ならびにターゲット症状にかかわる適切な用語の選定において有用である。

決定された検索ワードを用いて検索を行う際に重要となるのは「AND」と「OR」などの論理演算子の使用である。例えば“指標検査A”AND“ターゲット症状B”であれば「Aという指標検査を用いたBというターゲット症状名」の研究を検索することが可能となる。一方で，“指標検査A”OR“指標検査B”であれば「AまたはBという指標検査を用いた研究」を検索することができる。一つの指標検査またはターゲット症状に複数の表現があることは珍しくない。従って，指標検査およびターゲット症状の類義語・同義語を「OR」で接続して用いることで取りこぼしの少ない検索を実現し，「AND」を用いて不必要な文献の抽出を抑えることでより洗練化された検索を実施することが必要となる。なお，ターゲット症状の検索ではすでに定式化された“検索フィルター”を用いることも可能となっているが，現在その感度・特異度についての問題が指摘されている（Doust, Pietzak, Sanders, & Glasziou, 2005; Leeflang, Scholten, Rutjes, Reitsma, & Bossuyt, 2006; Ritchie, Glanville, & Lefebvre, 2007）。そのため，指標検査およびターゲット症状については，適切な用語の収集をもとに文献検索を行うことが推奨されている（de Vet et al., 2008）。なお，MEDLINEおよびEMBASEにおける検索フォーマットの例およびその詳細については，de Vet et al.（2008）を参照されたい。

レビューが科学的に質の高いものであるためには，文献検索においても再現性があるものでなければならない。検索プロセスの詳細な記述は結果の再現性を支えるだけにとどまらず，レビューに有用な情報源を示唆し，検索の厳密性・妥当性を担保するものとなる（Irwig, Tosteson, Gatsonis, Lau, Golditz, Chalmers, & Mosteller, 1994; Whiting, Rutjes, Dinners, Reitsma, Bossuyt, & Kleijnen, 2005; Mallett, Deeks, Halligan, Hopewell, Cornelius, & Altman, 2006; Moher, Tetzlaff, Tricco, Sampson, & Altman, 2007）。具体的な検索プロセスの記述について，コクラン共同計画では以下の内容の記載が求められている。(1)利用したデータベースまたは情報源（それらがカバーする文献の年代範囲），(2)用いた検索ワードまたは文献収集の方法（ハンドサーチおよびその他の検索方法），(3)検索を行った日時，(4)収集された文献の総数，(5)それぞれの電子データベースまたは収集方法で入手した文献の件数，(6)タイトルとアブストラクトから特定された目的論文の件数，(7)フルテキストで入手可能な論文件数，(8)参考文献リストにおける論文数，(9)最終的に分析の対象となった文献件数である。それに加えて，分析から除外された論文については，その理由を明記する必要がある。検索ワードおよび検索手続きはあらかじめ設定される必要があり，実行にあたっては検索日時を記録することが推奨されている。なお，これらの文献検索から適格・除外基準の適用を行う過程を論文中や付録において記載する際には，系統的レビューとメタアナリシスの報告ガイドラインであるPRISMA声明（Preferred Reporting Items for Systematic reviews and Meta-Analyses; Moher, Liberati, Tetzlaff, & Altman, 2009）のフローダイアグラムに従って報告することが望ましい（Figure 3）。

### 3.4 統合する研究の質の評価：QUADAS-2

診断精度研究のメタアナリシスでは，収集した一次研究の質の影響を大きく受ける（White et al., 2011）。一次研究の質とは研究のデザインや参加者の集め方，検査

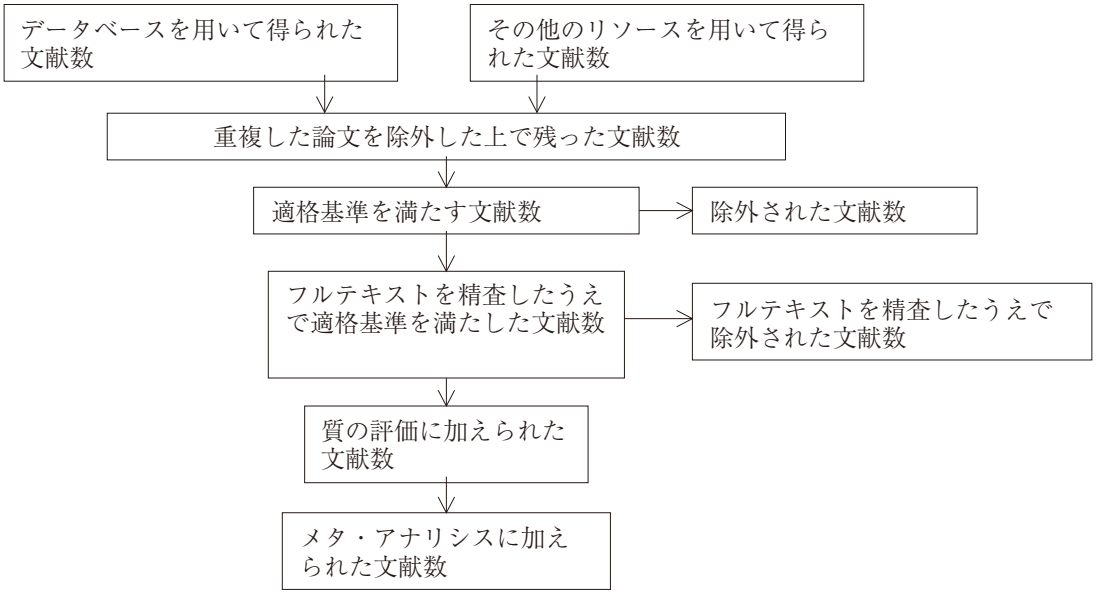


Figure 3　PRISMA における文献収集のプロセス（Moheret al.（2009）を元に作成）

の解釈の仕方などによって評価される（White et al., 2011）。レビューを質の高いものにし，診断精度について，できる限り正しい推定を行うためには，収集された一次研究の質の評価は欠かせないものとなる。また，一次研究の質を評価することによって，種々のバイアスの混入を排除するという目的もある。なお，種々のバイアスについては Table 5を参照されたい。「Handbook for DTA reviews」では一次研究の質を評価し，レビューに載せることを推奨している（Reistma, Whiting, Vlassov, Leeflang & Deeks, 2009）。またそこでは，QUADAS（Quality Assessment of Diagnostic Accuracy Studies; Whiting, Rutjes, Reitsma, Bossuyt, & Kleijnen, 2003; Whiting, Weswood, Rutjes, Reitsma, Bossuyt, & Kleijnen, 2006）という評価ツールを推奨している。

QUADAS は14項目によって構成されており，参加者の集め方，指標検査の解釈の仕方などについてどのように行われたかを「はい」「いいえ」「不明」の３択で評価する。すべての一次研究について14項目の評価を行い，それぞれの研究の質について項目の合計得点などをプロットする。「はい」の数が多いほど質の高い研究となり，各研究の質の高さや質のばらつきについて一目でわかるような図表の作成が望ましい。なお，「Handbook for DTA reviews」では QUADAS のうちの11項目を使用することが推奨されているが（Reistma et al., 2009），2011年に Whiting, Rutjes, Westwood, Mallett, Deeks, & Reitsma（2011）によって QUADAS の改訂版である QUADAS-２が作成されている（Table 6）。なお，RevMan では QUADAS-２に対応しており，今後 QUADAS-２が主流となることが予想される。そこで，本節では QUADAS-２の項目を元に一次研究の質の評価方法について述べる。

QUADAS-２は，11項目から構成されており，それらは参加者選択，指標検査，参照基準，フローとタイミングの四つの領域に分けられる。QUADAS-２を用いた一次研究の質の評価は，「レビュークエスチョンの明確化」，「レビューの目的に特化した調整」，「フローダイアグラムの作成」，「バイアスおよび適応可能性の評価」の４段階からなる。第 1 段階の「レビュークエスチョンの明確化」において，想定する母集団，指標検査，参照基準や対象となる症状について明確な定義を行う。第２段階の「レビューの目的に特化した調整」では，系統的レビューの目的に沿って，項目の追加・削除や採点方式について調整を行う。QUADAS-２の特定の項目はレビュー目的によっては不要かもしれないし，QUADAS-２の11項目だけではバイアスを評価しきれない可能性もある。次に，第３段階の「フローダイアグラムの作成」では，各一次研究がどのようなプロセスで参加者を集め，指標検査および参照基準を施行・解釈したのかフローダイアグラムを作成する。最後に，第４段階の「バイアスおよび適用可能性の評価」において，Table 6にある11項目を用いて各一次研究のバイアスおよび適用可能性について評価を行う。

Figure 4に評価結果の提示例を示した。QUADAS-２の各項目の重みづけや，総合的な評価を行うプロセスは

Table 5　診断精度研究におけるバイアス（Whiteet al., 2008；杉岡他，2014を元に作成）

| 領域 | バイアスの種類 | いつ起きるか | 診断精度に対する影響 | 対処方法 |
|---|---|---|---|---|
| 参加者 | スペクトラムバイアス | 参加者がターゲット症状を代表する集団ではない時。 | 疾患の範囲が対象となった参加者に限定される。 | 実践においてその検査を使うことが想定される集団をサンプリングする。 |
| | 選択バイアス | 参加者がランダムサンプリングされていない。 | 診断精度が過大評価される。 | 連続的あるいはランダムなサンプリングを行う。 |
| 指標検査 | 情報バイアス | 指標検査の解釈が参照基準の結果を知った上で行われる。あるいは，本来実践で得られる以上の情報を得て行われる。 | 診断精度が過大評価あるいは過小評価される。 | 指標検査の結果は本来実践において得られる情報のみで行う。 |
| 参照基準 | 分類不可バイアス | 参照基準によって参加者を陽性・陰性に正しく分類できていない。 | 診断精度が参照基準の精度に依存する。 | ターゲット症状について正しく分類する方法を参照基準とする。 |
| | 確認バイアス | 一部の参加者が参照基準を施行していない。 | 診断精度が過大評価される。 | すべての参加者に対して指標検査および参照基準の両方を施行する。 |
| | 組み込みバイアス | 指標検査が参照基準に組み込まれている。 | 診断精度が過大評価される。 | 指標検査および参照基準は独立なものにする。 |
| データ解析 | 除外データ | 検査結果が解釈不能だったケースを除外して分析する。 | 診断精度が過大評価される。 | すべての参加者が分析に組み込まれるべきであり，解釈不能だったケースを提示する。 |
| | 解釈バイアス | 指標検査および参照基準の結果を同一人物が解釈する。 | 診断精度が過大評価される。 | 指標検査および参照基準の結果を独立に解釈する。 |

Table 6　QUADAS-２の評価項目（Whiting, Rutjes, Westwood, Mallett, Deeks, & Reitsma, 2011を元に作成）

| | 項目 | 評価 |
|---|---|---|
| 参加者選択 | A. バイアスリスク | |
| | 参加者は連続あるいはランダムにサンプリングされたか | はい / いいえ / 不明 |
| | ケース・コントロール型研究ではないか。 | はい / いいえ / 不明 |
| | 不適切なデータの除外を行っていないか。 | はい / いいえ / 不明 |
| | 参加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 | |
| | 参加者の選択はレビュークエスチョンに合致しない可能性があるか。 | 懸念 : 低 / 高 / 不明 |
| 指標検査 | A. バイアスリスク | |
| | 指標検査の結果は参照基準の結果を知らない状態で解釈されたか。 | はい / いいえ / 不明 |
| | 閾値が用いられた場合，その閾値は事前に定義されていたか。 | はい / いいえ / 不明 |
| | 指標検査の実施および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 | |
| | 指標検査の実施および解釈はレビュークエスチョンに合致しない可能性があるか。 | 懸念 : 低 / 高 / 不明 |
| 参照基準 | A. バイアスリスク | |
| | 参照基準はターゲット症状を正しく分類していると仮定されるか。 | はい / いいえ / 不明 |
| | 参照基準の結果は指標検査の結果を知らない状態で解釈されたか。 | はい / いいえ / 不明 |
| | 参照基準の実施および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 | |
| | 参照基準により診断されたターゲット症状はレビュークエスチョンに合致しない可能性があるか。 | 懸念：低 / 高 / 不明 |
| フローとタイミング | A. バイアスリスク | |
| | 指標検査および参照基準の間に適切な期間が存在したか。 | はい / いいえ / 不明 |
| | すべての参加者に対し参照基準を施行したか。 | はい / いいえ / 不明 |
| | すべての参加者が同一の参照基準で分類されたか。 | はい / いいえ / 不明 |
| | すべての参加者が解析に含まれているか。 | はい / いいえ / 不明 |
| | 参加者のフローおよびタイミングによってバイアスが生じた可能性があるか。 | リスク : 低 / 高 / 不明 |

評価を行う前の段階で明確に定義されなければならない。あらかじめ決められたプロセスに従ってすべての一次研究について評価を行った後，その結果を表やグラフにまとめて分かりやすく提示する必要がある。一次研究の質を厳密に評価し，その手続きを明確にすることによってレビューの読者に対してレビューにおける問題提起やそのプロセスを明瞭に伝えることができる。また，一次研究の質が評価されレビューが行われることによって，以降に行われる一次研究全体の質向上にも寄与すると考えられる。

### 3.5　研究結果の統合方法

系統的レビューにおいては，最終的に収集した一次研究のデータを統合するメタアナリシスを行う。診断精度研究のメタアナリシスには，介入研究のメタアナリシスにはない難しさがある。診断精度研究のメタアナリシスでは，介入研究のように平均値差のような一つの指標ではなく，感度と特異度のような二つの指標を用いる必要がある。感度と特異度は，検査の閾値の設定によってトレードオフの関係で変化するものであり，そのような閾値の違いによる一次研究間のばらつき（閾値効果）も診断精度のメタアナリシスでは考慮する必要がある。そのため，単純にサンプルサイズによる重みづけ平均で検討することができない場合が多い。この点に関して，2000年代から，データ統合方法についてさまざまな研究がなされ，階層的なモデル（Reitsma, Glas, Rutjes, Scholten, Bossuyt, & Zwindeman, 2005; Rutter & Gatsonis, 2001）なども提案されてきている。本節では，Devillé, Buntinx, Bouter, Montori, de Vet, van der Windt, & Bezemer（2002），White et al.（2011），そして Macaskill et al.（2010）を参考にメタアナリシスの方法について述べる。

Devillé et al.（2002）によると，診断精度研究のメタアナリシスは，(1)個々の一次研究結果の提示，(2)異質性の検討，(3)閾値効果の検討，(4)異質性に対する対処，(5)統計的データ統合が適切ならどのモデルを使用するか決める，(6)統計的データ統合の実施の 6 段階からなるとされる。以下では，この 6 段階に従って，解説を行う。

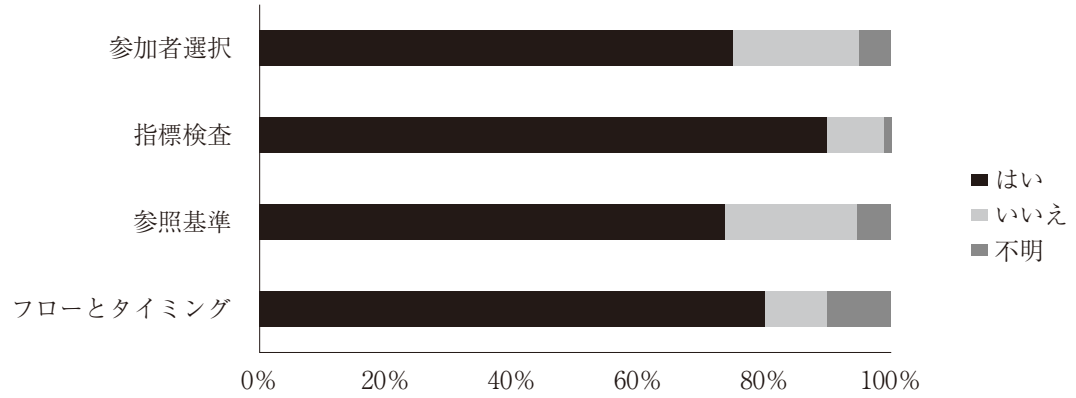


Figure 4　QUADAS-2 による評価結果の提示例

| | TP | FP | FN | TN | Sensitivity | Specificity |
|---|---|---|---|---|---|---|
| Kunisato et al.(2000) | 65 | 19 | 35 | 1 | 0.65(0.22, 0.75) | 0.95(0.70, 0.99) |
| Kunisato et al.(2014) | 95 | 28 | 5 | 42 | 0.95(0.90, 0.99) | 0.60(0.55, 0.65) |
| Sakamoto et al.(2007) | 14 | 5 | 6 | 95 | 0.70(0.55, 0.79) | 0.95(0.90, 0.99) |
| Sakamoto et al.(2010) | 10 | 2 | 0 | 8 | 0.50(0.25, 0.75) | 0.80(0.60, 0.85) |
| Somatori et al.(2011) | 32 | 6 | 8 | 14 | 0.80(0.10, 0.95) | 0.70(0.55, 0.79) |
| Somatori et al.(2012) | 25 | 20 | 25 | 80 | 0.50(0.25, 0.75) | 0.80(0.10, 0.85) |
| Suzuki et al.(2009) | 44 | 5 | 6 | 20 | 0.88(0.10, 0.99) | 0.80(0.10, 0.95) |
| Suzuki et al.(2014) | 27 | 8 | 3 | 12 | 0.90(0.70, 0.95) | 0.60(0.40, 0.70) |
| Tokita et al.(2005) | 30 | 6 | 20 | 44 | 0.60(0.55, 0.65) | 0.88(0.10, 0.99) |
| Tokita et al.(2008) | 16 | 24 | 24 | 36 | 0.40(0.25, 0.50) | 0.60(0.50, 0.70) |

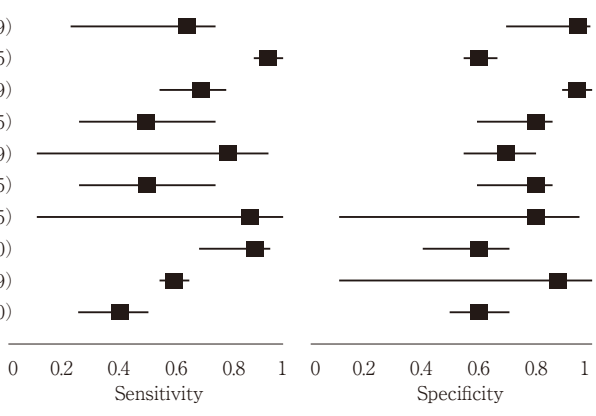


Figure 5　Forest plot（例）

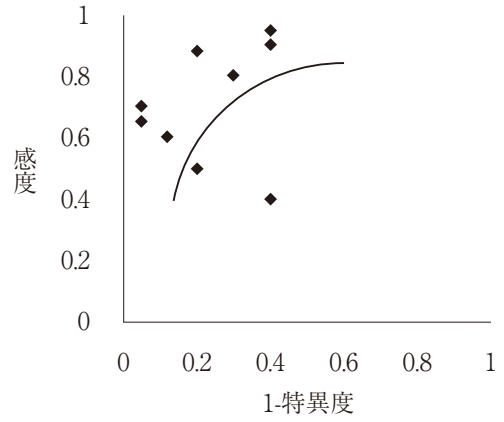


Figure 6　Summary ROC plot（例）

### 3.5.1　個々の一次研究結果の提示

一次研究を提示する主な方法として，forest plot と summary ROC plot（以下 SROC plot）がある。forest plot は，それぞれの一次研究の著者名，発刊年次，真陽性や偽陽性などの四つの指標，感度・特異度などについて記載し，感度と特異度とその信頼区間（もしくは標準誤差）をプロットしたものである（Figure 5参照）。これらの基本的な情報がプロットされていることによって，感度・特異度のばらつきの程度などについても，一目で理解することができる。なお，異質性にかかわる参加者の特徴や研究のデザインなどを付記する場合もある。また，forest plot には感度と特異度の二つの指標がプロットされることから，coupled forest plot と呼ぶこともある。

SROC plot は，ROC 平面上に個々の一次研究の感度・特異度をプロットしたものである（Figure 6）。RevMan の SROC plot では，個々の研究を四角形で表現する。その縦の長さは患者数に対する測定精度（つまり感度の精度），横の長さは非患者数に対する測定精度（つまり特異度の精度）を表し，大きいほど精度が高い（多くの場合，サンプルサイズも大きい）。SROC plot には，後に説明する統計モデルに基づいた SROC 曲線と要約感度・特異度（summary sensitivity and specificity point）を記載することも可能である。また，一次研究において複数の指標検査を実施して検査間比較を行うような研究もあり，系統的レビューにおいて検査間比較を行うこともある。その場合，一つの SROC plot 上に二つの検査の感度・特異度の値をどちらがどの検査かわかるようにプロットした上で，二つのプロットした点を実線でつないで，二つの検査結果が同一の研究から得られたものであることを示すことがある。このような SROC plot を Linked ROC plot と呼ぶ（Macaskill et al., 2010）。

### 3.5.2　異質性と閾値効果の検討と異質性への対処

異質性とは，一次研究における結果のばらつきの程度のことであり，一般的に診断精度研究のメタアナリシスにおいて異質性が確認されることが多い。Willis & Quigley（2011）によると，236の診断精度のメタアナリシスへの調査では，70%の研究において異質性が報告されている。異質性は，個々の一次研究における対象集団の違い，検査の実施方法，検査結果の解釈，参照基準の種類，閾値の違いやメタアナリシスに含めた研究のバイアスなどさまざまな要因の影響を受ける（White et al., 2011）。

異質性の評価は，forest plot や SROC plot を使用した視覚的な確認によって行うことが多い。forest plot を確認することで，一次研究における結果のばらつきを確認することができるが，そのばらつきが閾値の効果によるものかその他の要因によるものか判断することは難しい。そのため，まず閾値効果を検討する必要がある。

閾値効果を検討する方法としては，メタアナリシスに含んだ一次研究の感度と特異度間の Spearman の順位相関係数を算出する方法がある（Devillé et al., 2002）。一次研究の感度と特異度の間に強い負の相関がある場合，閾値効果が生じている可能性がある（Devillé et al., 2002）。また，SROC plot や統計モデルに基づいた SROC 曲線から閾値効果を検討することもできる。SROC 曲線が感度・特異度をプロットしたデータにフィットしている場合は閾値効果が生じていると考えられる。

SROC 曲線に沿った一次研究間のばらつきは閾値効果であり，一方その SROC 曲線から距離が遠くなるようなばらつきは，閾値効果の影響以外の要因で結果がばらついていると考える（White et al., 2011）。SROC plot によって，閾値効果とそれ以外の要因による異質性を検討することができる。なお，介入研究のメタアナリシスにおいては，異質性の指標として $I^2$ などを使用することが多いが，閾値効果を考慮できてないので，「Handbook for DTA reviews」では推奨されていない。また，後に紹介する階層モデルを用いて異質性の指標を推定することも可能ではあるが，その解釈が難しいという問題もある。

異質性への対処法として，Begg（2005）は，まず異

質性の原因を個々の一次研究にまで戻って調査すべきとしている。その上で取りうる方法としては，(1)適格・除外基準の変更やサブグループ解析，(2)ランダム効果モデルによる解析，(3)データ統合を行わず記述的に統合する，の三つがある。異質性の原因が明らかな場合は，適格・除外基準を厳しく設定するかサブグループ解析を行うことができる（Buntinx et al., 2009）。しかし，適格・除外基準を厳しくしたり，サブグループ解析をすることで，異質性を減らすことができるが，探索的に検討することは避ける必要がある。適格・除外基準の変更はデータ解析前に実施しその理由も報告する必要があり，サブグループ解析もプロトコルの段階で記載しておく必要がある（Buntinx et al., 2009）。また，「Handbook for DTA reviews」では，診断精度研究は異質性があるものとみなして，そもそもデータ統合においては異質性も考慮できるランダム効果モデルを用いるべきとしている（Macaskill et al., 2010）。最後に，異質性が強い場合は，無理にデータ統合を行わずに，記述的に統合することも必要となる（White et al., 2011）。

### 3.5.3 モデル選択

データ統合におけるモデルは，大きく分けて固定効果モデルとランダム効果モデルに分けることができる。固定効果モデルは，解析に含んだ個々の一次研究の結果に対して一つの真の値を仮定するものであり，個々の研究結果はランダム誤差によってばらつくと考える（White et al., 2011）。一方，ランダム効果モデルでは，個々の一次研究にそれぞれの真の値があり，個々の研究結果は真の値の違いおよびランダム誤差によってばらつくと仮定する。そのため，ランダム効果モデルでは，研究間の実際のばらつきである異質性を考慮にいれた解析ができる。「Handbook for DTA reviews」では，診断精度研究は異質性があるものとみなして，ランダム効果モデルを推奨しているが，研究間変動を推定できないほど一次研究の数が少ない場合や異質性や閾値効果がないと想定できる場合は固定効果モデルが適切としている（Macaskill et al., 2010）。また，Buntinx et al.（2009）のガイドラインでは，(1)異質性も閾値効果もない場合は，固定効果モデルによる感度と特異度の重みづけ平均，(2)異質性はないが閾値効果がある場合は，Moses-Litenberg 法（Moses, Shapiro, & Littenberg, 1993）による SROC 曲線，(3)異質性がある場合は，階層モデルを用いたランダム効果モデルによる解析を用いるとされる。以下では，診断精度研究のメタ分析に特化した Moses-Litenberg 法による SROC 曲線と階層モデルについて解説する。

### 3.5.4 Moses-Littenberg 法による SROC 曲線

Moses-Littenberg 法による SROC 曲線は診断精度研究におけるデータの統合方法として最も一般的な手法である（Macaskill et al., 2010）。Moses-Littenberg 法は固定効果モデルを採用して，異質性の指標がない，要約推定値，95％信頼区間，曲線下面積が正確でないなど，解析としての限界点がある。一方で SROC 曲線を用いた異質性の検討において利便性が高いため，RevMan では SROC plot に Moses-Littenberg による SROC 曲線が描出される。

Moses-Littenberg 法による SROC 曲線は，以下の方法で求められる。まず，個々の研究の感度と特異度から以下の D と S を求める。

$$D = logit（感度）- logit（1 - 特異度）\quad (10)$$

$$S = logit（感度）+ logit（1 - 特異度）\quad (11)$$

式(10)を展開すると D は，DOR の対数となり，式(11)を展開すると S は（TP*FP）／（FN*TN）の対数となる。S は検査陰性に対する検査陽性の割合であり，閾値の代理的な指標として使用することができる。各研究の D と S の値と式(12)を用いて，線形回帰モデルによって $\alpha$ と $\beta$ を求める。なお，その際，D を用いた分散逆数重みづけ法を用いることもある。

$$D = \alpha + \beta S + 誤差 \quad (12)$$

切片の $\alpha$ は DOR の対数の平均値と解釈する。$\alpha$ と $\beta$ が求められたら，最後に式(13)を用いて，SROC 曲線を引くことができる。

$$E(感度) = \frac{1}{1 + exp\left[\frac{\alpha + (1 + \beta)logit(1 - 特異度)}{1 - \beta}\right]} \quad (13)$$

Moses-Littenberg 法は，閾値の指標をモデルに組み込んでいるため，閾値効果のあるデータにも対応することができる。一方で，診断精度研究においては閾値以外にも多くの要因が個々の研究結果のばらつきに影響を与えている場合も多く（Macaskill et al., 2010），それらは Moses-Littenberg 法では検討できない。そのため，ランダム効果モデルを採用した階層的モデルも提案されている（Macaskill et al., 2010）。

### 3.5.5 階層的モデル

SROC plot などで異質性が確認された場合は，階層的モデルによるランダム効果モデルを用いたデータ統合を行い，より正確な診断精度の推定を行う必要がある。階

層的モデルには，Reitsma et al.（2005）による Bivariate モデルや Rutter & Gatsonis（2001）による Hierarchical SROC（以下 HSROC）モデルなどがある

Bivariate モデルでは，感度と特異度の平均値や分散，感度と特異度間の相関が推定できる。推定される感度と特異度の分散を利用することで異質性を検討することができる。Bivariate モデルは，ランダム効果モデルであり，レベル 1 において感度と特異度の研究内変動性が二項分布に従うとし，レベル 2 で logit 変換した感度と特異度の研究間変動性は正規分布に従うとしてモデリングする。

HSROC モデルでは，Moses-Littenberg 法と同じように $\alpha$ と $\beta$ を推定するのに加えて，閾値 $\theta$ も推定できる。また，$\alpha$ と $\theta$ の分散から異質性を検討することができる。HSROC モデルも，ランダム効果モデルであり，レベル 1 において感度と 1 －特異度の研究内変動性が二項分布に従うとする。また、診断精度を表す $\alpha$ と閾値を表す $\theta$ はランダム効果，ベータはそれらと独立した固定効果とする。レベル 2 で診断精度を表す $\alpha$ と閾値を表す $\theta$ は正規分布に従うとしてモデリングする。最終的に，$\alpha$ の平均値 A と $\beta$ から SROC 曲線を引くことができる。

階層的モデルは診断精度において予測される異質性の評価に長けており，これまでの固定効果モデルを仮定した統計手法よりも優れていると考えられている（White et al., 2011）。しかしながら，Willis & Quigley（2011）による，236の診断精度のメタアナリシスを調査した結果では，7 割の研究で感度と特異度の重みづけ平均や Moses-Littenberg 法が使用されており，Bivariate モデルや HSROC モデルの使用は少ないのが現状である。診断精度研究において異質性や閾値効果の問題は避けられないことを考えると，階層的モデルのさらなる普及が求められるといえる。

#### 3.5.6 検査間比較

診断精度研究の場合，複数の指標検査の結果の比較が研究目的となることがある。複数の指標検査結果を比較する方法には，⑴片方もしくは両方の検査を評価したすべての適格基準を満たした研究を利用する方法と，⑵同じ参加者が両方の検査を受けるもしくは参加者をランダム化してどちらかの検査を受けさせるような研究に限定する方法がある（Macaskill et al., 2010）。後者の方法の方ではバイアスが少ないが，そのような研究は少ないため，実施することが難しい場合が多い。前者の方法では，利用可能なすべての研究を分析に使うことができるという利点があるが，バイアスなどによって異質性が生じてしまう可能性がある。そのため，異質性も検討できる階層的モデルが良いとされる。しかしながら，そのように複数の診断精度を比較する一次研究がそもそも少ないため，必ずしもランダム効果モデルを仮定した方法が優れているとも限らない（Macaskill et al., 2010）。

### 3.6 結果の解釈と結論

結果の解釈を行う際，レビューの結果とそこから得られる知見を改めて記載する必要がある（Bossuyt, Davenport, Deeks, Hyde, Leeflang, & Scholten, 2013）。Bossuyt et al.（2013）では，解釈を記述する際に，考察として「主な結果の要約」,「レビューの長所および短所」，そして「レビュークエスチョンへの適用可能性」の三つの大枠に従って書くことが推奨されており，結論部分では「臨床的意義」と「今後の研究への示唆」の二つに触れることが求められている。以下では，Bossuyt et al.（2013）に従って，「Handbook for DTA reviews」において結果の解釈と考察において記載すべき内容を説明する。

#### 3.6.1 主な結果の要約

考察では，まずレビュークエスチョンを再掲する。次に Summary of Finding（SoF）表を作成し，分析によって得られた各数値や特記事項について簡潔かつ明瞭に記述する。なお，一次研究の質や異質性の評価結果も SoF 表に記載する。

SoF 表の主な目的は，結果の理解を促進することである。また，著者がレビューの解釈を結果のデータと照らし合わせ，データが解釈を支持していることを確認できるというメリットもある。なお，SoF 表は RevMan を用いて作成することができる。SoF 表では⑴表の上部に，レビュークエスチョンとその構成要素（母集団，セッティング，参照基準，指標検査）を記載する，⑵表の上部に，バイアスリスクの評価や適用可能性もしくは過度な異質性による限界点を記載する，⑶閾値が異なるなどで同じ指標検査が複数ある場合は，それぞれごとに分けて記載する，⑷メタアナリシスによる診断精度の推定に含めた一次研究の数と参加者数を記載し，診断精度を感度・特異度で表し，推定値にかかわる統計的不確実性を信頼区間などで記載し，有病率も記載する，⑸複数の指標検査の比較を行った際はそれぞれの診断精度を求めるために用いた一次研究の数やそれぞれの診断精度の推定値，そして統計的な検定結果などを記載する。

診断精度のメタアナリシスで用いられるような複雑な統計手法は多くの読者に浸透しているとは言えない。そのため，統計の専門用語を用いることは最小限にして，得られたデータを数値または文章で再掲することによって読者の理解を促進する必要がある。

#### 3.6.2 レビューの長所および短所

系統的レビューの結果の解釈において，レビューの質の評価が必要になる。系統的レビューの質の評価については，(1)レビューに含めた一次研究の質と(2)レビュー過程の質の二種類がある。これらを評価して，今回のレビューの長所と短所について述べる。まず，系統的レビューに含めた一次研究の特徴，質，量，結果の一貫性など要約する。そして，一次研究の限界については，QUADAS-2の4領域（参加者選択，指標検査，参照基準，フローとタイミング）をそれぞれ参照してまとめる(Bossuyt et al., 2013)。

次に，レビュー過程での長所と短所について述べる。レビュー過程には，文献検索や適格・除外基準，質の評価やデータ抽出，そして分析において限界が存在する。文献検索の限界点としては，どんな検索フィルターを使用しても，潜在的にバイアスの原因になる可能性があり，対処が必要になる。質の評価やデータ抽出における限界点としては，一次研究において適切な報告がなされてない場合に，質の評価や抽出ができないことが挙げられる。分析における限界点としては，一般にメタアナリシスの推定精度は個々の研究よりも高くなるが，レビューに入れた研究数が少なく異質性が高い場合は精度が悪くなる点などがある。

#### 3.6.3 レビュークエスチョンへの適用可能性

レビュー結果の適用可能性について考察することは，診断精度のメタアナリシスを行うにあたって特に重要である。ランダム効果モデルによる診断精度の推定値は，含まれた全研究における平均推定値である。診断精度の推定には，研究のデザイン，対象とする集団，指標検査，参照基準などが影響を与えうるため，適用可能性を評価することが必要になる。

収集された一次研究は，レビュークエスチョンに応じた母集団や研究デザインのものであることが望ましい。また，QUADAS-2ではレビュークエスチョンに関連する項目も含まれているため，これらを参照しながらメタアナリシスの知見をどれだけレビュークエスチョンに適用することができるか考察する必要がある。

#### 3.6.4 臨床的意義

レビューで得た知見の範囲を超えず，データによって結論を正当化しつつ，できるだけ実用的かつ明確に書く。臨床的意義を考察するとき，クリニカルパスにおける指標検査の位置づけや，指標検査は意図した役割（ほかの検査に追加する，ほかの検査に代わる，スクリーニングに使う）をどのくらい果たしているのか，指標検査が陽性・陰性を示した場合の結果についても考慮されるべきである。

#### 3.6.5 今後の研究への示唆

診断検査を臨床場面に適用する上で，診断の精度以外に必要な追加の研究について言及し，その具体的な研究デザインや方法についても記載する。また，今回メタアナリシスによって得られた知見のみでは診断精度を評価できない場合，報告の質の高い研究など具体的にどのような研究が今後望まれるかについても述べる。

## 4. おわりに

ここまで，主にコクラン共同計画が作成している「Handbook for DTA reviews」を中心に診断精度のメタアナリシスの手法について述べてきた。診断精度研究の系統的レビューとメタアナリシスの方法に関するガイドラインについては，少しずつまとまってきている。実証に基づく臨床実践を行う上で，診断やアセスメントに関するエビデンスは，その他の発症要因や予後，治療効果などのすべてのエビデンスの基礎となるものである。また，疾患の検査方法についてのエビデンスの蓄積は，クライエントの健康状態を素早く正確に特定し，その後の治療を行う上では不可欠である。そのため，介入効果のエビデンスの統合だけではなく，診断精度に関するエビデンスの統合も非常に有益であると考えられる。

本稿で参考にした論文の多くが医学系の文献であった。しかしながら，診断精度に関する事柄は，実践において質問紙や投映法，神経心理学的検査などさまざまな検査を使用する臨床心理学においても無視できない。今後は，臨床心理学においても，種々の心理検査の診断精度が検討され，臨床実践に役立てられるようになることが期待される。診断精度のエビデンス統合はいまだ発展途上の領域ではある。しかしながら，実践家および研究者がその方法論について十分な知識を得ておくことは有益であると考えられる。

## 引用文献


Begg, C. B. (2005). Systematic reviews of diagnostic accuracy studies require study by study examination: first for heterogeneity, and then for sources of heterogeneity. *Journal of Clinical Epidemiology*, **58**, 865–866.

Bossuyt, P. M., Davenport, C., Deeks, J. J., Hyde, C., Leeflang, M. M., & Scholten, R. (2013). Chapter 11: Interpreting results and drawing conclusions. In J. J. Deeks, P. M. Bossuyt, C. Gatsonis (ed.), *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy* Version 0.9. The Cochrane Collaboration. <Available from: http://srdta.cochrane.org/.> (December 13, 2013)

Bossuyt, P. M., Leeflang, M. M. (2008). Chapter 6: Developing Criteria for Including Studies. *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version* 0.4 The Cochrane Collaboration. <Available from: http://srdta.cochrane.org/.> (September, 2008)

Buntinx, F., Aertgeerts, B., & Macaskill, P. (2009). Guidlines for conducting systematic reviews of studies evaluating the accuracy of diagnostic tests Knottnerus, A, J., Buntinx, F (Edit) *The Evidence Base of Clinical Diagnosis: Theory and Methods of Diagnostic Reasearch.* 2 nd edition Hoboken, NJ: Wiley-Blackwell 180–212.

de Vet, H. C. W., Eisinga, A., Riphagen, I. I., Aertgeerts B., & Pewsner, D. (2008). Chapter 7: Searching for Studies. *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy* Version 0.4 The Cochrane Collaboration. <Available from: http://srdta.cochrane.org/.> (September, 2008)

Deeks, J. J., Wisniewski, S., & Davenport, C. (2013). Chapter 4: Guide to the contents of a Cochrane Diagnostic Test Accuracy Protocol. In J. J Deeks, P. M. Bossuyt, C. Gatsonis (ed.), *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy* Version 1.0.0. The Cochrane Collaboration. <Available from: http://srdta.cochrane.org/.> (September 13, 2013)

Devillé, W. L., Buntinx, F., Bouter, L. M., Montori, V. M., de Vet, H. C. W., van der Windt, D. A., & Bezemer, P. D. (2002). Conducting systematic reviews of diagnostic studies: didactic guidelines. *BMC Medical Research Methodology*, **2**, 9.

Doust, J. A., Pietrzak,E., Sanders, S., Glasziou, P.P. (2005). Identifying studies for systematic reviews of diagnostic tests was difficult due to the poor sensitivity and precision of methodologic filters and the lack of information in the abstract. *Journal of Clinical Epidemiology*, **58**, 444–449.

Fraser, C., Mowatt, G., Siddiqui, R., Burr, J. (2006). Searching for diagnostic test accuracy studies: an application to screening for open angle glaucoma (OAG). *XIV Cochrane Colloquium*, **88**, 23–26. (abstract)

古川壽亮 (2000). エビデンス精神医療：EBP の基礎から臨床まで 医学書院

Irwig, L., Tosteson, A.N., Gatsonis, C., Lau J., Colditz, G., Chalmers T.C., Mosteller, F. (1994). Guidelines for meta-analyses evaluating diagnostic tests. *Annals of Internal Medicine*, **120**, 667–76.

Leeflang, M. M., Scholten, R. J., Rutjes, A.W., Reitsma, J.B., & Bossuyt, P. M. (2006). Use of methodological search filters to identify diagnostic accuracy studies can lead to the omission of relevant studies. *Journal of Clinical Epidemiology*, **59**, 234–40.

Macaskill, P., Gatsonis, C., Deeks, J. J., Harbord, R. M., & Takwoingi, Y. (2010). Chapter 10: Analysing and Presenting Results. In J. J. Deeks, P. M. Bossuyt, C. Gatsonis (ed.), *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy* Version 1.0. The Cochrane Collaboration. <Available from: http://srdta.cochrane.org/.> (December 23, 2010)

Mallett, S., Deeks, J. J., Halligan, S., Hopewell, S., Cornelius, V., Altman, D. G. (2006). Systematic reviews of diagnostic tests in cancer: review of methods and reporting. *BMJ*, **333**, 413.

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Annals of internal medicine*, **151**, 264–269.

Moher, D., Tetzlaff, J., Tricco, A. C., Sampson, M., & Altman, D. G. (2007). Epidemiology and reporting characteristics of systematic reviews. *PLoS medicine*, **4**, 78.

Moses, L. E., Shapiro, D., & Littenberg, B. (1993). Combining independent studies of a diagnostic test into a summary ROC curve: Data-analytic approaches and some additional considerations. *Statistics in Medicine,* **12**, 1293–1316.

Reitsma, J. B., Glas, A. S., Rutjes, A. W., Scholten, R. J., Bossuyt, P. M., & Zwinderman, A. H. (2005). Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *Journal of clinical epidemiology*, **58**, 982–990.

Reistma, J. B., Whiting, P., Vlassov, V., Leeflang, M. M., & Deeks, J. J. (2009).

Chapter 9: Assessing methodological quality. In J. J. Deeks, C. Gatsonis (ed.) *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy* Version 1.0.0. The Cochrane Collaboration. <Available from: http://srdta.cochrane.org/.> (October 27, 2009)

Ritchie, G., Glanville, J., Lefebvre, C. (2007). Do published search filters to identify diagnostic test accuracy studies perform adequately? *Health Information and Libraries Journal*, **24**, 188–92.

Rutter, C. M., & Gatsonis, C. A. (2001). A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Statistics in Medicine*, **20**, 2865-2884.

Sackett, D. L., Rosenberg, W., Gray, J. A., Haynes, R. B., & Richardson, W. S. (1996). Evidence based medicine: what it is and what it isn't. *BMJ*, **312**, 71-72.

Smith, B. J., Darzins, P. J., Quinn, M., & Heller, R. F. (1992). Modern methods of searching the medical literature. *The Medical Journal of Australia,* **157**, 603-611.

Song, F., Khan, K. S., Dinnes, J., & Sutton, A. J. (2002). Asymmetric funnel plots and publication bias in meta-analyses of diagnostic accuracy. *International journal of epidemiology*, **31**, 88-95.

杉岡隆・野口善令・大西良浩 (2014). 診断法を評価する～いつも行っている検査は有効か？～ 特定非営利活動法人健康医療評価研究機構

丹野義彦 (2001). エビデンス臨床心理学 －認知行動理論の最前線－ 日本評論社

Virgili, G., Conti, A., Murro, V., Gensini, G. & Gusinu, R. (2009). Systematic reviews of diagnostic test accuracy and the Cochrane Collaboration. *Internal Emerg Medicine*, **4**, 255-258.

White, S., Schultz, T., & Enuameh, Y. A. K. (2011). Synthesizing evidence of diagnostic accuracy. *Lippincott Wiliams & Wilkins*.

Whiting, P., Rutjes, A. W., Dinnes, J., Reitsma, J. B., Bossuyt, P. M., & Kleijnen, J. (2005). A systematic review finds that diagnostic reviews fail to incorporate quality despite available tools. *Journal of clinical epidemiology*, **58**, 1-12.

Whiting, P., Rutjes, A., Reitsma, J. B., Bossuyt, P. M., & Kleijnen, J. (2003). The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Medical Research Methodology*, **3**, 1-13.

Whiting, P., Rutjes, A. W., Westwood, M. E., Mallett, S., Deeks, J. J., Reitsma, J. B., & Bossuyt, P. M. (2011). QUADAS-2 : a revised tool for the quality assessment of diagnostic accuracy studies. *Annals of internal medicine*, **155**, 529-536.

Whiting, P., Westwood, M., Burke, M., Sterne, J., Glanville, J. (2008). Systematic reviews of test accuracy should search a range of databases to identify primary studies. *Journal of Clinical Epidemiology*, **61**, 357-364.

Whiting, P. Westwood, M., Rutjes, A., Reitsma, J., Bossuyt, P., & Kleijnen, J. (2006). Evaluation of QUADAS, a tool for the quality assessment of diagnostic accuracy studies. *BMC Medical Research Methodology* **6**, 1-8.

Willis, B. H., & Quigley, M. (2011). The assessment of the quality of reporting of meta-analyses in diagnostic research: a systematic review. *BMC Medical Research Methodology*, **11**, 163.