Beginners' Training Sheet for Systematic Review

ver.6.1 by last updated on June 29, 2018
南郷　栄秀　Eishu NANGO, MD, PhD
http://spell.umin.jp

採用論文　　Reviewer:　　年　月　日

authors：

title：

citation：

PubMed PMID：

quick check list

1. 論文の PICO は何か？
2. コクランレビューか？
3. GRADE approach を用いているか？
4. 全ての研究を網羅的に集めようと努力したか？
   ①検索に用いた文献データベースは何か？
   ②どのような検索語を用いたか？
   ③どの期間の研究を調べたか？
   ④どのような種類の研究を調べたか？
   ⑤個々の論文の参考文献まで追跡して調べたか？
   ⑥個々の研究者や専門家に連絡を取ったか？
   ⑦出版されていない研究も探したか？
   ⑧英語以外で書かれた研究も探したか？
5. 全ての研究が網羅的に集められたか？
6. 集められた研究の risk of bias は評価されたか？
   ①複数の評価者によって評価されたか？
   ②どのような評価基準で評価されたか？
7. 結果の評価

論文の構造

要約 abstract, summary
緒言 introduction
方法 methods　　←チェックすべき項目はほとんどここにある！
結果 results
考察 discussion

研究が扱っている題材は，「要約 abstract，summary」に記載されており，これは PICO でまとめることができる．ただし，「要約 abstract，summary」の部分だけでは情報が不十分なことが多く，論文の「方法 methods」の項で詳細を確認することが必要である．

0 システマティックレビューという研究手法

0-1）システマティックレビューとは？

システマティックレビューSystematic review; SR(系統的レビュー)は，特定の疑問に関して，数多くの研究を網羅的に再現性のある方法に従って集め，その時点における結果のまとめを行ったもの（図 1 の b と c）である．多くは，ランダム化比較試験 Randomized controlled trial; RCT をまとめたものであるが，コホート研究や診断の研究を統合したものもある．このシートでは，RCT をまとめたシステマティックレビューの読み方について解説する．

定義：明確に定式化された疑問について，関連する研究を特定し，選択し，批判的に吟味し，レビューに組み入れられた研究から得られるデータをまとめて解析する，系統的で明示的な方法を用いて行うレビューのこと（Cochrane library, 1998）．**論文には，タイトルにシステマティックレビューSystematic review やメタアナリシスMeta-analysis と書かれていることが多い．**

目的：対象となる疑問についての現在利用可能な全ての研究データをまとめること

特徴：①問題とするトピックに関する研究を網羅的に探し出す膨大な努力をしていること
②各論文を批判的に評価していること
③事前に定められた質評価基準を満たす研究のみを統合して結論が導かれていること

一方，システマティックレビューでない一般のレビュー（総説，図 1 の a）は**ナラティブレビューNarrative review** と呼ばれ，著者の主観的な意見を述べたものにすぎない．ナラティブレビューの多くは専門家によって書かれたものであるが，あらゆる研究が網羅されているわけではなく，著者の意見を支持する先行研究の結果だけを都合良く引用している可能性が高く，引用された研究の質もバラバラである．時にシステマティックレビューとは異なる結論が導き出されることもある．

0-2）メタアナリシス Meta-analysis とは？

前述のように，システマティックレビューは，研究結果を“体系的に集める”ことに重きが置かれている．これに対して，メタアナリシス Meta-analysis とは，この用語を最初に使用した Glass（1976）の定義によれば，“研究結果を統合する目的で，個々の研究から得られた解析結果の膨大なコレクションに対して実施する統計解析”である．すなわち，集められた複数の研究の結果を，統計学的手法を用いて統合したものをメタアナリシスと呼ぶ（図 1 の c と d）．

メタアナリシスが行われないシステマティックレビュー（図 1 の b）では，個々の研究結果を羅列するのみにとどまったり，統計学的解析を行うことなく著者の直感で結論が導き出されたりする．これに対して，メタアナリシスでは，一定の統計学的手法で統合するため，誰が行っても同じ結論を得ることができる．しかし，統合の手法が優れていても，手当たり次第に研究結果を集めてきたのでは，統合された結果にはバイアスが含まれてしまう（図 1 の d）．したがって，システマティックレビューであり，かつメタアナリシスが行われているものが，結論を信頼することができる（図 1 の c）．

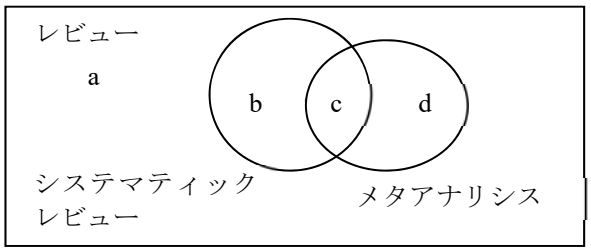


図 1：システマティックレビューとメタアナリシスの関係

0-3）システマティックレビューの批判的吟味とは？

システマティックレビューの批判的吟味は，このシステマティックレビューの質を評価することである．しかし質の評価には，

①システマティックレビューが正しく行われたのか（網羅的に研究が集められたのか）
②システマティックレビューの中で採用された研究の質の評価（risk of bias）
③統合された結果（メタアナライシスで計算された数字など）の質（エビデンスの質・確実性・強さ）

などがあり，混乱しやすいので注意して区別する必要がある．システマティックレビューそのものの質（作られ方が正しいか）が高くても，採用された研究が質の低いものばかりで，統合された結果の質も低い（結果が信頼できない）ことがある．

1 論文の PICO は何か？

PICO とは疑問を定式化したものであり，どんな患者が（P: Patient），どんな治療や検査を受けるのは（I: Intervention），何と比べて（C: Comparison），どうなるか（O: Outcome）を一文にまとめたものである．

患者 P(Patient)

介入 I(Intervention)：

比較 C(Comparison)：

結果 O(Outcome)：

システマティックレビューには複数の研究結果が集められており，個々の研究の PICO（年齢や病期などの患者背景，投与薬剤の用量など）は互いに異なることが多い．したがって，**システマティックレビューの批判的吟味で PICO を探る際は，おおよその内容を把握するだけでよい．**

よく分からなければ，導入 Introduction の一番最後をみると良い．この研究で検証しようとしている研究仮説に書いて簡潔に書かれている．おおよその PICO を把握するためのキーワードは，“the aim of this study”である．

PICO は 3 つあることに注意したい．①自分の臨床の疑問の PICO，②システマティックレビューの著者らが集めようとした PICO，③実際に集まった研究の PICO．もし，②と③が異なっていれば，このシステマティックレビューの統合した結果の質が低くなってしまう（**非直接性 indirectness**）．

2 コクランレビューか？

コクランレビューは，英国に本部がある世界的組織「コクラン Cochrane」（http://www.cochrane.org/）が作成するシステマティックレビューである．「コクラン」は複数の臨床試験の結果を総括的に評価する方法についての研究・開発を進めており，日々システマティックレビューを量産している． コクランが作成するシステマティックレビューは「コクランレビューCochrane Database of Systematic Review: CDSR」と呼ばれており，厳密な作成基準に従って作成され，査読されているため，コクランレビューであれば比較的質が高く信頼できる．コクランレビューはコクランライブラリ Cochrane Library: http://www.thecochranelibrary.com/（有料）にデータベース化されている．

□コクランレビューである
□コクランレビューでない

2018 年現在，コクランレビューの表紙は，以下のようなデザインである．

2014 年に「コクラン」の日本支部であるコクランジャパン（http://square.umin.ac.jp/cochranejp/）が設立され，2017 年には NPO 法人となった．コクランジャパンでは，コクラン系統的レビュー作成のサポートおよびトレーニングを提供し，日本の医療や政策の科学的根拠に基づいた意思決定を促進している．

コクランレビューの著者がコクランレビューでない SR を作成していることがあり，その場合も質の高い SR であることが期待できる．著者がコクランレビューを作成しているかどうかは，コクランライブラリで著者名を検索するとわかる．

3 GRADE approach を用いているか？

GRADE approach ／GRADE system とは，エビデンスの質と推奨の強さをグレーディングするための一般的かつわかりやすい方法としてカナダの McMaster 大学の Gordon Guyatt と Holger Schünemann らの GRADE working group（http://www.gradeworkinggroup.org/index.htm）が中心となり開発されたものである．**現在システマティックレビューと診療ガイドラインの国際標準作成方法として広く用いられている**．「GRADE」とは，「Grading of Recommendations Assessment, Development and Evaluation」の略である．

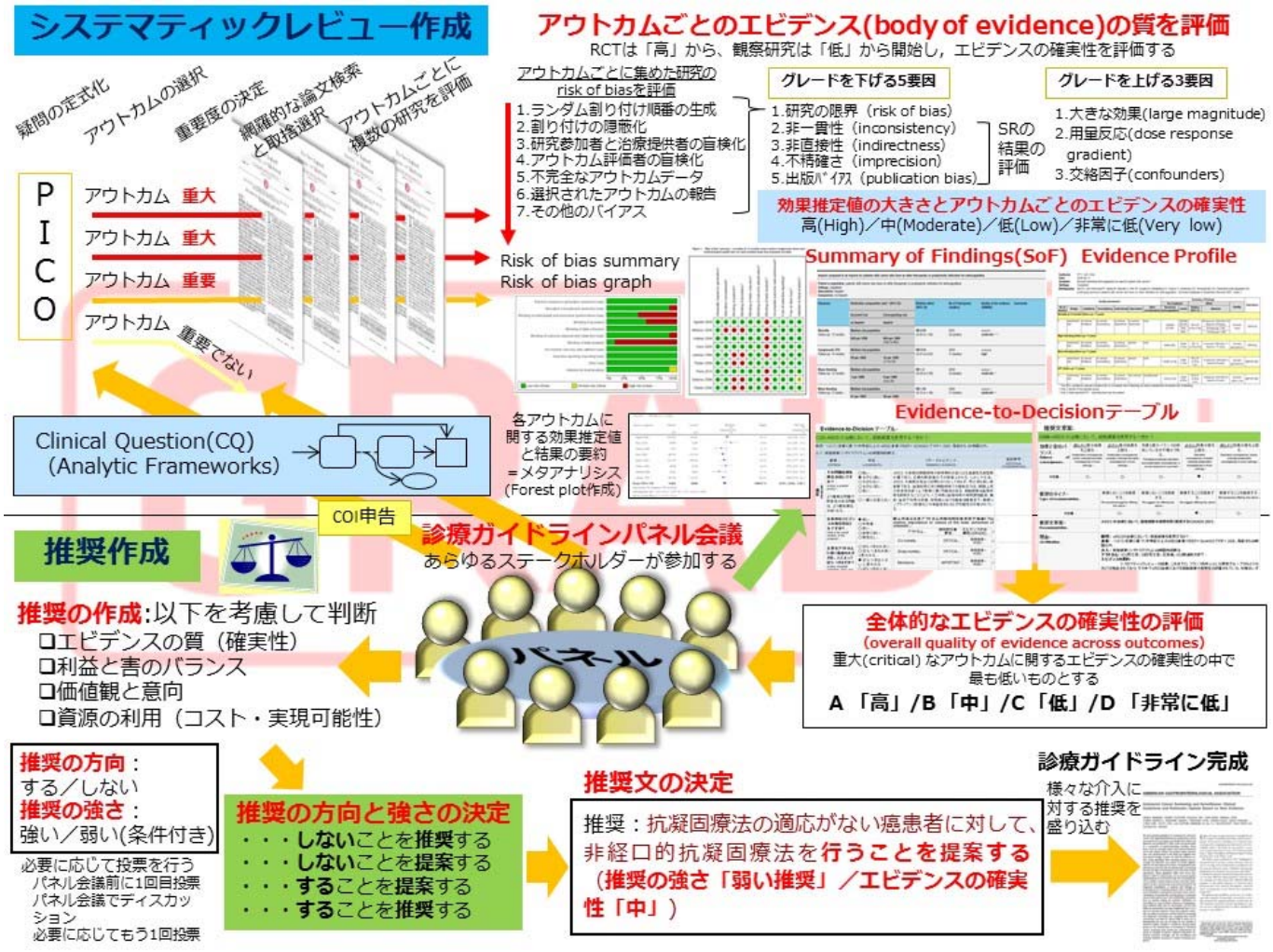
☐GRADE approach を用いて作られている
☐GRADE approach は用いられていない

GRADE approach を用いて作られているか否かの判断は容易ではない．論文中に「GRADE」の文字があっても，必ずしも GRADE approach で作られたものとは限らないからである．**Summary of Finding (SoF) table（12 ページ）があれば，GRADE approach で作られている可能性が高い．**

GRADE approach の特徴は以下の通りである．

- 関連するエビデンスを網羅的に収集する（＝システマティックレビュー）
- アウトカムごとにエビデンスを評価する（アウトカム中心主義）
- エビデンス総体（アウトカムごとに集められたエビデンスの一群）の質（＝効果推定値の確実性）を評価する（エビデンスの質は，エビデンスの確実性，エビデンスの強さ、などと表記されることもある）
- エビデンスの質の評価は，研究デザインをもとに，バイアスのリスク，非一貫性，非直接性，不精確さ，出版バイアスの 5 つの grade down の要因と，大きな効果，用量反応勾配，交絡因子の 3 つの grade up の要因を検討して，最終的に 4 段階の質（高，中，低，非常に低）に等級付けする
- 診療ガイドラインで推奨を決める際に，エビデンスの質，利益と効果のバランス，患者の価値観，コストやリソースをバランスよく検討する
- 診療ガイドラインの推奨は，「推奨の方向」（その医療行為を行う／行わない），「推奨の強さ」（強い／弱い），「エビデンスの質（確実性）」（高，中，低，非常に低の 4 段階）の 3 要素からなる

GRADE system の全体像



4 全ての研究を網羅的に集めようと努力したか？

研究はどのようにして集められたか？（研究を集めるフローチャートがあれば参照する）
①検索に用いた文献データベースは何か？
- □MEDLINE　□CINAHL
- □EMBASE　□ISI Web of Science
- □CENTRAL（Cochrane Central Register of Controlled Trials）／Cochrane Library
- □Google scholar
- □その他のデータベース（　　）

②どのような検索語を用いたか？（　　）
③どの期間の研究を調べたか？（　　）
④どのような種類の研究を調べたか？
- □ランダム化比較試験（RCT），準ランダム化比較試験（quasi-RCT），臨床対照試験（non-RCT，CCT）
- □システマティックレビュー（SR），メタアナリシス（MA），コクランレビュー（CDSR）
- □コホート研究
- □症例対照研究
- □診断研究（横断研究）
- □その他の種類の研究（　　）

⑤個々の論文の参考文献まで追跡して調べたか？
- □参考文献まで調べた
- □参考文献は調べなかった，不明

⑥個々の研究者や専門家に連絡を取ったか？
- □連絡を取った
- □連絡を取らなかった，不明

⑦出版されていない研究も探したか？
- □探した
- □探さなかった，不明

⑧英語以外で書かれた研究も探したか？
- □探した
- □探さなかった，不明

研究を集める際には，通常以下の流れに従う．
①研究テーマである PICO からキーワードを抜き出し，それを用いてデータベースを検索する．
②①を補完する形で参考文献を調べたり，専門家に直接連絡を取ったりして，データベースに含まれていない研究や未発表のデータを探す．
③集めた研究のうち，質の低い研究や他の研究とデータが重複しているもの，PICO が異なるものなどを除外する．

システマティックレビューでは，研究の集め方のルールとして inclusion criteria と exclusion criteria を定めていることがある（これは個々の研究で研究参加者を集めるために定めている inclusion / exclusion criteria とは異なるので混同しないこと）．右の図のようなフローチャート（PRISMA flow diagram）が本文中にあれば，それを参考にすると良い．この研究では，キーワードを用いてデータベースから論文を検索した後，RCT でないものを 209 件，重複報告を 70 件，現在進行中の試験を 5 件除外して固有の RCT のみ残し，さらに exclusion criteria に該当するものを除外して，最終的に 13 件をレビューの対象としている．

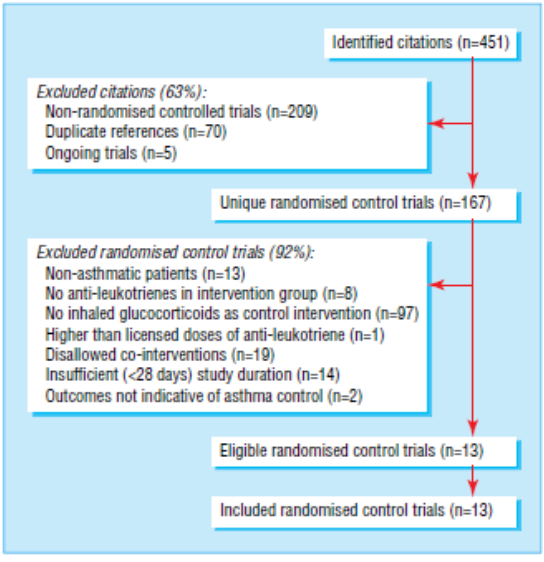


Fig 1 Selection process of eligible randomised controlled trials from all identified citations

研究の検索の手順は Methods の冒頭に記載がある．以下の点をチェックする．

①検索に用いた文献データベースは何か？:主として Medline だが，これだけでは不十分である．他に EMBASE, CINAHL, Cochrane Library, Web of Science, Google Scholar などがある．Cochrane Library には，CDSR（Cochrane Database of Systematic Review），CENTRAL（Cochrane Library に収載されている RCT のデータベース）がある．さらに個々の専門領域のデータベースなども併用されるべきである．

②どのような検索語を用いたか？：検索範囲は広すぎても狭すぎても良くない．紙面の都合上，キーワードのみが記載されている場合がある．その際，検索語は Supplement や Appendix に掲載されている．

③どの期間の研究を調べたか？ ：検索日が古いと，より新しい研究は含まれていない可能性が高くなる．

④どのような種類の研究を探したか？：テーマのカテゴリーによって集める対象の研究は異なる．

⑤個々の論文の参考文献を追跡したか？:データベースにはない研究が参考文献にあるかも知れない．“screened bibliographies”とか“search reference literature”などと記載されていることが多い．

⑥研究者や専門家に連絡を取ったか？：論文には書かれていないデータも集めたか？“contact to researchers”や“contact to authorities”などと記載されていることが多い．

⑦出版されていない研究も探したか？：治療効果が認められなかった場合など，インパクトの弱い研究結果は発表されにくい傾向にある．ClinicalTrials.gov（http://clinicaltrials.gov/）などの臨床試験の登録データベースが参照されることもある（→次項ファンネルプロットを参照）．

⑧英語以外で書かれた研究も探したか？：言語バイアス language bias を排除する．例えば，結果に差がなかった論文はインパクトが低いので英語で発表されず，日本語で発表されているかもしれない．英語の文献のみを抽出したものは，効果が過大評価されていると考えるべきである．“no language restrictions”などと記載されていることが多い．

5 全ての研究が網羅的に集められたか？

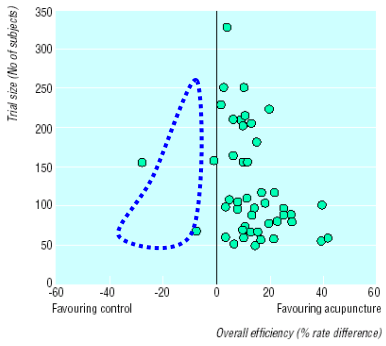
□研究数が 9 件以下である　→明らかな出版バイアスがあるとは言えない
□研究数が 10 件以上である
**　○ファンネルプロットを用いて出版バイアスの有無を検討している**
**　　□ファンネルプロットは対称　　→出版バイアスはない**
**　　□ファンネルプロットは非対称　→出版バイアスがある**
**　○ファンネルプロットは用いられていない**
**　　□出版バイアスは明らか**
**　　□出版バイアスはなさそう**
**　　□不明**

研究を網羅的に集めようとすることと，実際に洩れなく集めることができたかどうかは，別の問題である．網羅的に研究が集められたか，本来存在するはずの研究が抜けていないか，すなわち**出版バイアス publication bias** が検討されているか確認する．記載箇所を探すキーワードは，“funnel plot”，“publication bias”，“Egger’ regression test”（連続変数で用いる），“Begg’s test”（順序変数で用いる）である．出版バイアスの評価には研究数が 10 件以上（5 件以上という人もいる）必要である．

ファンネルプロット funnel plot

ファンネルプロット funnel plot は，SR において出版バイアスの存在を評価するために用いられる．右図の例に示すように，ファンネルプロットは通常，横軸に試験の治療効果（例えばオッズ比や有効率の差）を，縦軸には治療効果の精度 precision（多くは症例数や分散によって規定される）を示し，この 2 次元上に個々の臨床試験の結果がプロットされる．

プロットした結果は，理想的には左右対称の正規分布になるはずである．しかし，一般的に研究から得られた治療効果が小さいほど，その研究結果が発表されにくくなる．このような出版バイアスが存在する場合，ファンネルプロットの下方の片側は空白となる．本来そこには介入を無効とする少数例の研究の結果が存在するはずであるが，これらが発表されていないために，出版済みの研究のみが統合された結果は，真の効果よりも高い評価となってしまう．ファンネルプロットの図から，統合された結果が有効か無効のどちらの方に傾くかを予想しておくことが重要である（右図では左下が欠けているので，集められた研究を統合した結果は真の効果よりも効果がある方に傾いている）．ファンネルプロットは，論文中に示されていなくても，個々の論文のデータをもとに自分で作成することができる．なお，出版バイアス以外にファンネルプロットで左右が対象とならない原因としては，研究の質のバラツキがある．



ファンネルプロット（BMJ 1999;319:160 より）

6 集められた研究の risk of bias は評価されたか？

①複数の評価者によって評価されたか？
□複数の評価者　→何人で評価されたか？
→評価者間で評価のくい違いが生じた場合
○合意を形成して最終的に評価を下している
○合意を形成せず，各評価者の判断を個別に記載している
○その他
□単独の評価者
□不明

②どのような評価基準で評価されたか？
□Cochrane risk of bias tool で評価した
□Jadad score で評価した
□それ以外の評価基準で評価した
→どのような評価か？
介入研究の場合の評価するべき項目
○ランダム割付け順番の生成 random sequence generation
○割付け方法の隠蔽化 allocation concealment
○研究参加者と治療提供者のマスキング blinding of participants and personnel
○アウトカム評価者のマスキング blinding of outcome assessment
○不完全アウトカムデータ incomplete outcome data
○選択されたアウトカムの報告 selective outcome reporting
○その他のバイアス other sources of bias
□明確な基準はない

評価者の数は Methods の欄に記載がある．
集められた研究を単独の評価者が評価すると，研究の妥当性の評価に偏りが出る可能性がある．このため，集められた研究の評価は，**複数の評価者で独立に**行うことが望ましい．評価者間で評価にくい違いが生じた場合は，話し合いや 3 番目の評価者により最終的な判断が下されることが多い．
論文評価の基準は Methods の欄に記載がある．
集められた研究は，明確な基準を持って評価される必要がある．いわゆる原著論文の批判的吟味である．
かつては，RCT の質を評価するスコアのうち唯一妥当性が証明されている Jadad score（ハダッド・スコア）[8]と呼ばれるものが用いられていた．ここでは個別の研究の質であり，SR そのものやその結果の質でない．

Jadad score[0)]
1．その研究はランダム割付けと明示されているか？（1＝はい，0＝いいえ）
2．ランダム割付けの方法について明示されていて，かつ適切か？（0＝明示されていない，1＝明示されていてかつ適切），－1＝明示されているが不適切）
3．その研究はダブルブラインドと明示されているか？（1＝はい，0＝いいえ）
4．ダブルブラインドの方法について明示されていて，かつ適切か？（0＝明示されていない，1＝明示されていてかつ適切），－1＝明示されているが不適切）
5．試験の最初と最後の各治療群の患者数を明らかにするために，投与中止 withdrawals や脱落 dropout が記載されているか？（1＝はい，0＝いいえ）
0～5 点で表記．概ね 3 点以上であれば，比較的質が高いと判定される．

しかし，必ずしも各ドメインが同じ重みにはならないのに合計点数を算出して総合評価していたり，チェック項目についての記載をされていない方が，不適切に記載されている場合よりもスコアがよくなったりすることから，評価の妥当性に疑問があり，最近では用いられなくなってきている．現在では，**コクランが作成した Cochrane Handbook for Systematic Reviews of Interventions[9)]にある「Cochrane Collaboration risk of bias tool」が用いられる**ことが多い．GRADE system もこのコクランの **risk of bias（RoB）**評価ツールを用いている．

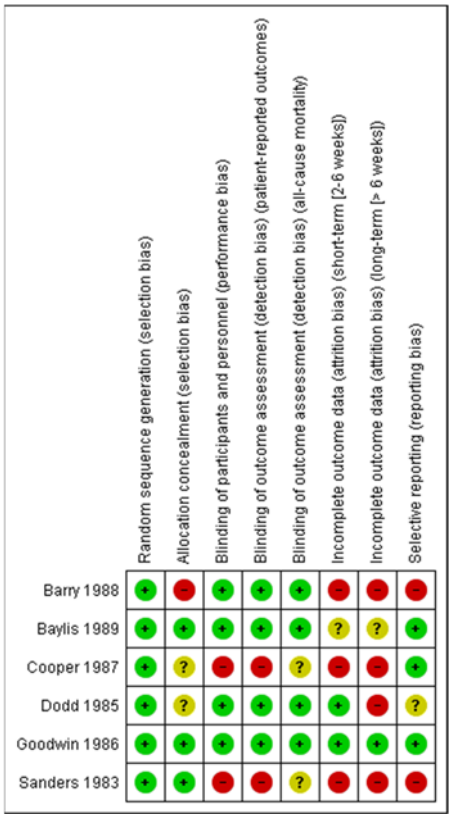
Cochrane handbook のランダム化比較試験バイアスのリスク risk of bias 評価ツール [9)]

①ランダム割り付け順番の生成 random sequence generation（選択バイアス selection bias）
②割り付けの隠蔽化 allocation concealment（選択バイアス selection bias）
③研究参加者と治療提供者のマスキング blinding of participants and personnel（施行バイアス performance bias）
④アウトカム評価者のマスキング blinding of outcome assessment（検出バイアス detection bias）
⑤不完全なアウトカムデータ incomplete outcome data（症例減少バイアス attrition bias）
⑥選択されたアウトカムの報告 selective outcome reporting（報告バイアス reporting bias）
⑦その他のバイアス other sources of bias

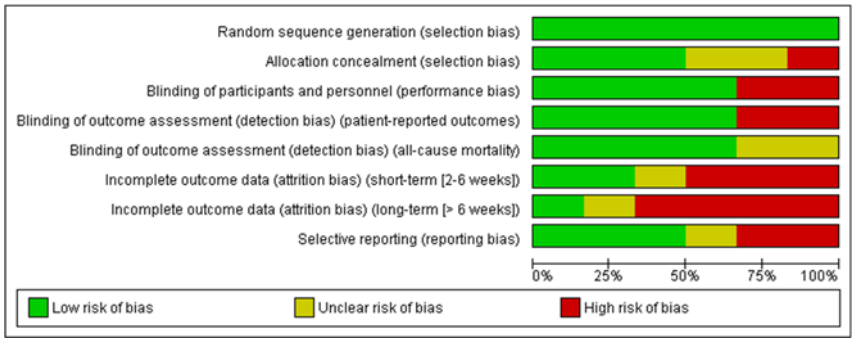
Risk of bias ツールはスケールやチェックリストではない．ドメインに基づいた評価を行う．すなわち，各ドメインは独立して評価する．また，マスキングの有無はアウトカムによって影響が異なるので，本来は risk of bias の評価はアウトカムごとに行うべきであるが，コクランレビューを含む多くのシステマティックレビューでは，全てのアウトカムを通して 1 つの評価のみされている．

コクランの risk of bias 評価ツールを用いて評価した結果は，以下のような Risk of bias summary や Risk of bias graph で示される．

Risk of bias summary



Risk of bias graph



（Cochrane handbook より）

RCT 以外の研究デザインのシステマティックレビューでは，チェック項目も異なる．

観察研究の risk of bias 評価ツールには，RoBANS（Risk of Bias Assessment Tool for Nonrandomized Studies）というものがある（J Clin Epidemiol 2013;66:408）．GRADE アプローチでは，以下の基準を用いている

バイアスのリスク Risk of bias	説明（以下に問題があれば，high risk of bias とする）
①適切な適格基準を確立していない，あるいは適用していない（対照群の組み入れ）	● 症例対照研究におけるマッチングが過少または過大になっている ● コホート研究において，曝露したと曝露していない人が背景の異なる集団から選ばれている
②曝露およびアウトカムの双方における測定の不備	● 曝露やアウトカムの測定が不確かな場合（例：症例対照研究における思い出しバイアス） ● コホート研究において，曝露群と非曝露群で曝露内容やアウトカムの調査方法が異なっている
③交絡が十分にコントロールされていない	● コホート研究において，すべての既知の予後因子を測定してない，もしくは正確に測定していない ● 曝露群と非曝露群で予後因子や背景因子が一致していない，または解析の際にそれ他の統計学的な調整（例：多変量解析・傾向スコア）がされていない
④追跡が不十分または観察期間が短すぎる	● 特に前向きコホート研究において，アウトカムが発症するのに十分な観察期間がとられていない ● 曝露群と非曝露群の観察期間が異なっている

7 結果の評価

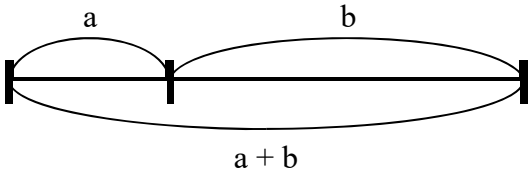
結果はどのように示されたか？フォレストプロットが示されていれば，これを読む．
□票数計算 vote counting
□リスク比 risk ratio（**相対危険度** relative risk）
□オッズ比 odds ratio
□リスク差 risk difference
□平均値の差 mean difference
□その他

1）結果の示し方

票数計算 vote counting は，効果ありの研究の数と効果なしの研究の数を単純に数えて比較したもの（有効：無効＝○：△のようなもの）である．この場合，いずれの研究の質も結果も，重みが等しいとみなされている．票数計算では，その介入が実際にはどの程度の効果を示すのかが分からないので，この方法で効果の有無を表現している場合には注意が必要である．ただ，集まった研究の PICO が違いすぎてメタアナリシスできないような場合には，統合せずに結果を一覧表にして表記することもある．

2）オッズ比とリスク比

Meta-analysis で統合される Outcome 指標には，リスク比，オッズ比，リスク差，平均値の差などがあるが，それぞれの 95%信頼区間 95%CI や p 値 p-value も評価すべきである．



	Outcome		
	(＋)	(－)	
介入群	a	b	a+b
対照群	c	d	c+d
	a+c	b+d	a+b+c+d

右上の表で，outcome 発生率と，介入群と対照群の outcome 発生率の比であるリスク比は，以下のように計算される．

介入群の Outcome 発生率 $= \frac{a}{a+b}$　　対照群の Outcome 発生率 $= \frac{c}{c+d}$

リスク比 Risk ratio; RR $= \frac{a(c+d)}{(a+b)c}$

一方，オッズは率と似て非なるものであり，ある事象が起こったものと起こらなかったものの比である．また，介入群のオッズと対照群のオッズの比をオッズ比といい，以下のように計算される．

介入群の Outcome 発生オッズ $= \frac{a}{b}$　　対照群の Outcome 発生オッズ $= \frac{c}{d}$

オッズ比 Odds ratio; OR $= \frac{ad}{bc}$

リスク比とオッズ比の式を見比べると分かるように，a＋b を b と，c＋d を c と置き換えることができれば，両者の値は一致する．a や c が小さい場合，つまり有病割合の低い疾患ほどオッズ比はリスク比に近似できる．

原則として，リスク比 risk ratio: RR を用いる．これは解釈が容易だからである．ただ，RR には対称性の性質がない（介入に対するアウトカム発症の RR の逆数をとってもアウトカム非発症の RR にならない）ので，ネットワークメタアナリシスのような場合は対象性の性質のあるオッズ比 odds ratio: OR を用いる．治療群と対照群の効果に差が無ければ，RR，OR は 1 に等しくなり，治療に効果がある場合は RR と OR は共に 1 未満，効果がない場合には RR，OR は 1 を越える． また，95%信頼区間 95%CI が 1 をまたいでいるかどうかによって，その結果が統計学的に有意かどうかが判断できる．対数オッズ logarithmic odds で示される場合には，ゼロ点から両側に距離の等しい対数スケールにプロットされる．

オッズ比と患者イベント予想発生率（PEER）から NNT が計算できる（はじめてトライアルシート参照）．この患者イベント予想発生率は，個々の患者でどれくらいの高さになるかを見積もるものである．

$$\mathrm{NNT} = \frac{1-\{\mathrm{PEER}\times(1-\mathrm{OR})\}}{(1-\mathrm{PEER})\times\mathrm{PEER}\times(1-\mathrm{OR})}$$

3）フォレストプロット forest plot

フォレストプロット forest plot は，Meta-analysis に特有のグラフである．またの名を**串刺し図，ブロボグラム brobogram** ともいう．個々の研究の結果は 1 行ずつならべられ，最下段にこれらを統合した結果が示される．この図から，各研究の結果が均一であるかがチェックできる．

フォレストプロットの表し方は研究グループによって異なるが，英国のグループでは，下図のように平均値を四角で示し，95%信頼区間を線で表している．四角の大きさは結果の精度で決まる（症例数の多い研究ほど四角が大きい傾向がある）．四角が大きく，信頼区間が短ければより信頼のおける研究結果といえる．

個々の研究を統合した結果は，菱形で示される．菱形の中心が代表値であり，菱形の大きさは，95%信頼区間を示す．すなわち，菱形を見て，それが 1（対数オッズでは 0）をまたいでいなければ，有意差ありという結論になる．

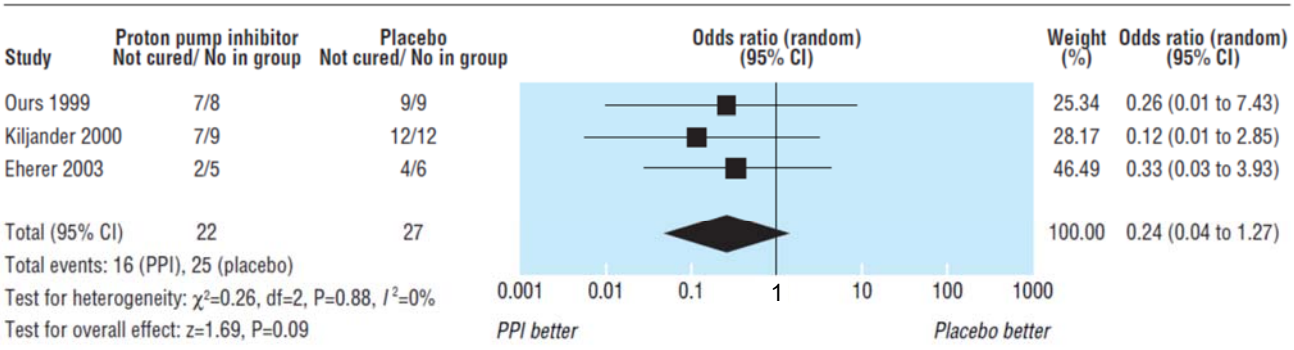
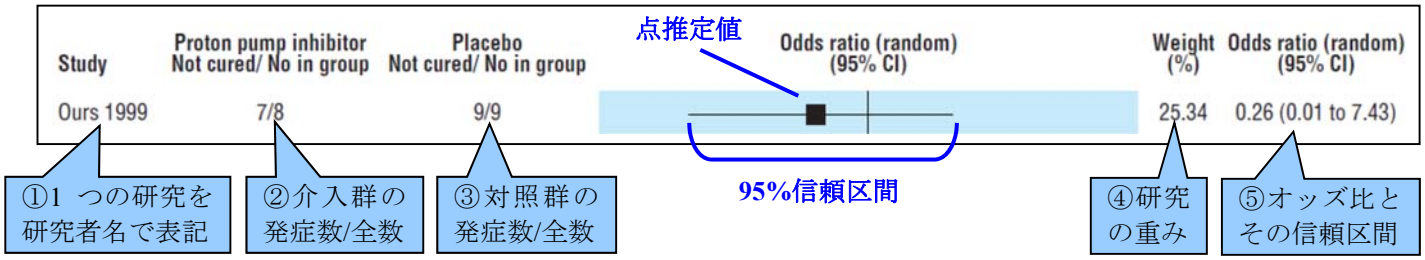


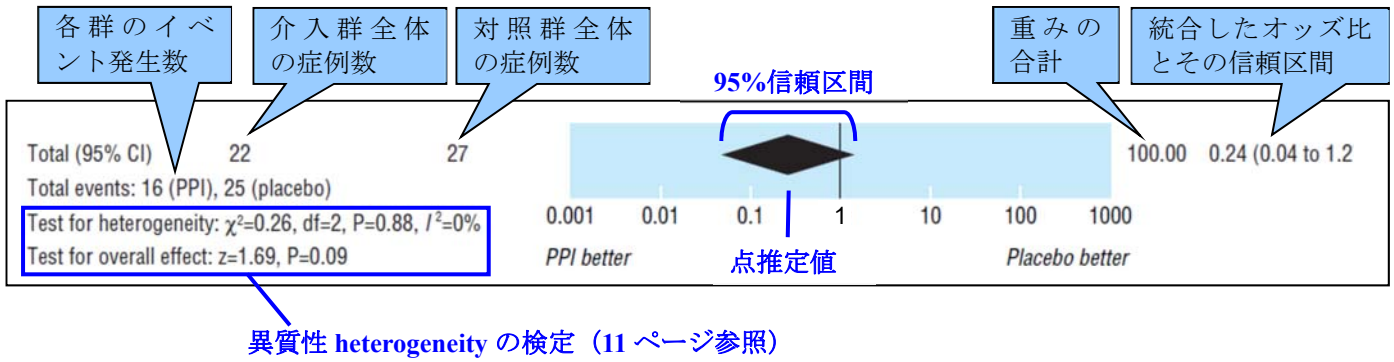
Fig 1 Meta-analysis of primary outcome (clinical failures—that is, patients still had cough at the end of the trial or reporting period), analyses by intention to treat (49 participants included in meta-analysis)

forest plot の例：Systematic review and meta-analysis of randomised controlled trials of gastro-oesophageal reflux interventions for chronic cough associated with gastro-oesophageal reflux. *BMJ* 2006;332:11-7 より

一例として，1 行目の研究について見てみよう．左から順に，その研究の筆頭著者（first author）の名前①が示され，続いて介入群と対照群それぞれの発症数／症例数②③，さらに各研究の重み付け（weight）④と並び，フォレストプロットには，点推定値と 95%信頼区間が四角と線で示されている．図の右側には，点推定値と 95%信頼区間が実数⑤で表記されている．この 1 行で 1 つの研究結果が表現されていることになる．縦に走る 0 の線と 95%信頼区間の線が縦に走る 1 の線と交叉している研究は，有意差がなかったことを示す．



最下段にある全ての研究の統合結果を見てみよう．菱形の中央が統合された効果の点推定値にあたり，幅が 95%信頼区間となる．研究間の異質性についての結果も forest plot に記載されている場合があるが，ない場合には本文中に記載がある．



4）結果の統合 meta-analysis

複数の研究結果について統計学的手法を用いて統合し，1 つの総合的な統計量で示したものをメタアナリシス Meta-analysis; MA という．統計学的手法で統合する際には，**Fixed effect model** や **Random effects model** が用いられる．集められた研究結果のバラツキはもっぱら偶然誤差であると仮定するのが Fixed effect model であり，この偶然によるバラツキ以外に，研究間に，プロトコールの違い，患者の違い，地域の違いなどといった無視できない違い（heterogeneity）としての各研究の偏りも関与していると考えるのが Random effects model である．Random effects model の方が信頼区間の幅が若干広く，有意差が出にくくなる．

具体的な統計手法としては，Outcome 指標の種類に応じて，Fixed effect model では，Mantel-Haenszel method, Peto method，General variance-based method が，Random effects model では DerSimonian-Laird method が用いられる．Fixed effect model と Random effects model のどちらが用いられるべきかについては議論が多く，どちらが優れているということはない．異質性 heterogeneity が低いと判定された場合には，いずれの方法を用いても同じ結論に達するので，どの方法が用いられているかは重要ではなくなる．しかし，異質性が高い場合は，Random effects model の方が結果が信頼できる．

異質性が極めて高い場合は，結果を統合しないという決断が下されることもある．あるいは，patient を性別，重症度，年齢など，サブグループに分けることで異質性がなくなる場合には，サブグループ解析 subgroup analysis を行って統合することができる場合もある．元の研究にバイアスが含まれている可能性が高い場合や，元の研究同士が著しく異なる場合は，統合してはいけない．

5）異質性 heterogeneity

異質性の検定についての記載は Methods の Statistical analysis の欄に記載がある．

「同質」とは，集められた全ての研究にバラツキがないことを指す．**異質性 heterogeneity** は統計学的異質性 statistical heterogeneity ともいい，研究の PICO のバラツキである臨床的異質性 clinical heterogeneity（臨床的多様性 clinical diversity）と，研究手法の質にバラツキをもたらす方法論的異質性 methodological heterogeneity（方法論的多様性 methodological diversity）の両方が含まれる．

各研究における効果の大きさが大きくバラついており，例えば信頼区間が重ならない場合は異質性があると考えられる．このため，**Cochran Q 統計量**と呼ばれる各研究間の結果のバラツキの度合いを計算し，これが有意なものかどうか，**カイ二乗検定**を用いて統計学的に検討する．ここでは，異質性の検討がなされたかだけを確認し，異質性の検定の結果については「6）結果の評価の仕方」の項で確認する．

6）結果の評価の仕方

メタアナリシスされている場合は，フォレストプロット forest plot を中心に結果の評価を行う．

PICO 毎に，統合された結果を評価する

→フォレストプロット毎の PICO を確認する（タイトルと横軸を見ると分かりやすい）

①採用された研究の種類，数，症例数は？

②統合された結果に有意差があるか？

③統合された結果の大きさ（点推定値と信頼区間）は？

④集められた研究に異質性 heterogeneity はあるか？

□異質性は検討されていない

□異質性は検討されている

異質性 heterogeneity の検定

○Cochran Q（カイ二乗検定）

□有意差なし　→同質

□有意差あり　→異質

○I^2 統計量

□0～25%　→異質性が低い

□25～50%　→異質性が中程度

□50～75%　→異質性が高い

□75～100%　→異質性が極めて高い

⑤異質性が高い場合には，高い原因は何か？

○患者背景が異なる

○介入内容が異なる

○アウトカムの定義が異なる

○研究の質が異なる

⑥異質性の高いものに対してサブ解析，感度分析が行われていれば，その結果を評価する

結果は PICO 毎に評価する．特に，全体を統合した結果と，Patient，Intervention/Comparison，Outcome の内容によってサブグループ解析が行われることが一般的なので，どれで有意な差が出ているかを確認する．

異質性は，「5）異質性 heterogeneity」の項に示したとおり，**Cochran Q 統計量**と**カイ二乗検定**を用いて統計学的に検討する．カイ二乗検定の結果はフォレストプロット（結果の図）の下の方に記載されていることが多い（8 ページ参照）．但し，Cochran Q 統計量のカイ二乗検定では検出力（power）が小さいため，集められた研究数が少ないときには有意水準（異質か同質かの判定基準）には 5%ではなくて 10%が用いられることもある（すなわち p<0.10 で有意差あり）．最近では，この Cochran Q の欠点を補うため **I^2 統計量**が用いられる．これは Cochran Q から自由度を引いたものを 100%表示したもので，負の値は 0%と見なすことから，I^2 統計量は 0～100%の値の間を取る． 25%以下では異質性が低く，25～50%は中程度，50～75%は高く，75～100%は極めて高いと評価する．これにより，統合する研究数が少なくても多くても，比較的正しく評価できる．実際のフォレストプロットでは，Cochran Q 統計量のカイ二乗検定と I^2 統計量の両方が併記されていることが多い．

異質性が高い場合には，研究間のバラツキがあることを示しているので，メタアナリシスされた結果は信頼性が低くなる．そのため，異質性が高い結果では PICO の違い（例えば，患者の重症度や薬剤の量）で分けてメタアナリシスし直したり（サブ解析），フォレストプロットで他のものから外れた研究を排除してメタアナリシスし直したり（感度分析）して，異質性が低くなるか検討する．サブ解析や感度分析により異質性が低くなれば，その結果は信頼性が高くなる．

以上をまとめると，メタアナリシスの結果を評価するための手順としては，①研究の種類，数，症例数を見て，フォレストプロットの統合された結果（一番下にある菱形）を見て，②有意差の有無と③効果の大きさを評価し，その後，④異質性の評価のために Cochran Q のカイ二乗検定に有意差があるかどうか，また I^2 統計量を確認し，異質性が高いと判断されるなら，⑤その原因は何かを，PICO の違いや研究の質のバラツキを見て検討する．⑥異質性の高いものに対してサブ解析，感度分析が行われていれば，その結果を評価する．

7）Summary of Finding(SoF) table

SR の結果を表にまとめる．この表には，アウトカムごとに介入群と対照群のリスクと相対効果，患者数および研究数，エビデンスの質（GRADE）が示されている．相対効果はフォレストプロットで示されているメタアナリシスの結果である．エビデンスの質は，集められた研究が統合された結果の質であり，GRADE approach によって 5 つのグレードダウンの要因と 3 つのグレードアップの要因での評価によって調整された後の各アウトカムの効果推定値の信頼性を示している．

LMWH compared with UFH for perioperative thromboprophylaxis in patients with cancer

Patient or population: patients with perioperative thromboprophylaxis in patients with cancer
Settings: inpatient
Intervention: LMWH
Comparison: UFH

Outcomes	Illustrative comparative risks* (95% CI)		Relative effect (95% CI)	No of participants (studies)	Quality of the evidence (GRADE)	Comments
	Assumed risk UFH	Corresponding risk LMWH				
Mortality Follow-up: median 2 weeks	42 per 1000	37 per 1000 (31 to 45)	RR 0.89 (0.74 to 1.08)	9938 (9 studies)	⊕⊕⊕○ moderate[1]	
PE Follow-up: median 2 weeks	6 per 1000	5 per 1000 (2 to 10)	RR 0.73 (0.34 to 1.54)	5825 (13 studies)	⊕⊕⊕○ moderate[1]	
DVT (symptomatic) Follow-up: median 2 weeks	9 per 1000	4 per 1000 (2 to 11)	RR 0.5 (0.2 to 1.28)	3233 (8 studies)	⊕⊕⊕○ moderate[1]	
Major bleeding Follow-up: median 2 weeks	47 per 1000	40 per 1000 (25 to 65)	RR 0.85 (0.52 to 1.37)	3533 (8 studies)	⊕⊕⊕○ moderate[1]	
Wound hematoma Follow-up: median 2 weeks	105 per 1000	72 per 1000 (55 to 93)	RR 0.68 (0.52 to 0.88)	2442 (6 studies)	⊕⊕⊕○ moderate[2]	
Thrombocytopenia Follow-up: median 2 weeks	11 per 1000	15 per 1000 (6 to 33)	RR 1.33 (0.59 to 3)	1911 (4 studies)	⊕⊕○○ low[1,2]	

*The basis for the **assumed risk** (e.g. the median control group risk across studies) is provided in footnotes. The **corresponding risk** (and its 95% confidence interval) is based on the assumed risk in the comparison group and the **relative effect** of the intervention (and its 95% CI).
CI: confidence interval; **DVT:** deep venous thrombosis; **LMWH:** low molecular weight heparin; **PE:** pulmonary embolism; **RR:** risk ratio; **UFH:** unfractionated heparin.

GRADE Working Group grades of evidence
High quality: Further research is very unlikely to change our confidence in the estimate of effect.
Moderate quality: Further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate.
Low quality: Further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate.
Very low quality: We are very uncertain about the estimate.

[1] The 95% CI includes both negligible effect and important benefit or important harm.
[2] Possible selective outcome reporting as few of the 16 included studies report on this outcome.

（*CDSR*2014:CD009447 より）

参考文献

1） Oxman AD, Cook DJ, Guyatt GH, for the Evidence-Based Medicine Working Group: Users' guides to the Medical Literature. IV: How to Use an overview? JAMA 1994;272:1367-1371.
2）開原成允，浅井泰博，治療や予防に関する文献の使い方，JAMA 医学文献の読み方，中山書店 2001 年，95-109.
3） Sackett DL et al. Evidence-Based Medicine, How to Practice and Teach EBM. Churchill Livingstone 2000, 133-138.
4） Bedenoch D 他著，斉尾武郎監訳，EBM の道具箱．中山書店 2002 年，33-38.
5)柳川敏彦，Ian Roberts，津谷喜一郎．META とは何かー未公表臨床試験を探す「アムネスティ」の試みー http://www.sphere.ad.jp/cont/27_3/META/report.html
6） Moher D, Cook DJ, Eastwood S, et. al. Improving the quality of reports of meta-analysis of randomised controlled trials: the QUOROM statement. The Lancet 1999;354:1896-900.
7） 丹後俊郎，メタアナリシス入門，エビデンスの統合を目指す統計手法，朝倉書店 2002 年.
8) Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJ, Gavaghan DJ, and others. Assessing the quality of reports of randomized clinical trials: is blinding necessary? Control Clin Trials 1996;17(1):1-12.
9）Higgins J, Green S. Cochrane Handbook for Systematic Reviews of Interventions 5.1.0. http://www.cochrane-handbook.org/.
10)森實敏夫，わかりやすい医学統計学，メタアナリシス（4），統計学的手法（1），Medical Tribune 2000 年，http://www.medical-tribune.co.jp/mtbackno3/3305/05hp/M3305281.htm.

改訂履歴

1.0→1.1
・書き込み用 CAT sheet を作成
1.1→2.0
・用語を統一
・結果の評価の項の説明を変更
2.0→2.1（2004.3.21）
・書き込み用 CAT sheet の改変（1 ページ化）
・論文の PECO を探るの項を改変
2.1→3.0（2006.3.10）
・全体の構成を大幅に改訂
・Narrative review との違いについて言及
・Jadad score についての解説を追加
・統計学的手法について言及
・ファンネル・プロットの記載場所を変更
・リスク比とオッズ比についての解説を追加
・フォレストプロットの解説を図示
3.0→3.1（2006.7.23）
・メタアナリシスの説明を充実化
・細かいデザインの変更
3.1→3.2（2006.9.23）
・複数の評価者による評価の項目に選択肢を追加
・heterogeniety の説明を改変
3.2→3.3（2008.1.28）
・複数報告されている研究の有無についてのチェック項目を新設
・集められた研究の結果は統合されたか（Meta-analysis）？の項の記載を訂正
・ブロボグラム brobogram の用語を追加
3.3→3.4（2009.2.20）
・一部の項目で選択肢の順序を変更
・研究が網羅的に集められたかの判定の参考になる文章を追加
・集められた研究の妥当性評価と Jadad score の解説を改正
・異質性の検定の評価方法についての解説を追加
3.4→3.5（2010.4.24）
・ファンネル・プロットの図の評価の仕方についての解説を追加
・集められた研究の妥当性評価の説明を修正
・異質性の評価の項目の並びを変更
・異質性の項に Cochran Q 統計量と I^2 検定の記述を追加
・研究の結果の統合の項に，統合してはいけない場合の記述を追加
・forest plot の解説に用いる例を差し替え
3.5→3.6（2010.10.3）
・誤字訂正
3.6→4.0（2010.11.22）
・PECO を PICO に変更
・「システマティック・レビュー」を「システマティックレビュー」に，「メタ・アナリシス」を「メタアナリシス」に，「ファンネル・プロット」を「ファンネルプロット」それぞれ変更
・全ての研究を網羅的に集めようと努力したか？の項に，どのような種類の研究を調べたか？を新たに追加．また，検索に用いた文献データベースに Google scholar を追加
4.0→5.0（2012.3.20）
・どのような種類の研究を調べたか？の項の選択肢を修正
・結果は統合されたか（Meta-analysis）？の項に，最終的に何件の研究が残り，採用されたか，の質問を追加
・異質性の検討のうち，結果の部分を「結果の評価」に移動
5.0→5.1（2012.3.25）
・誤字訂正
5.1→6.0（2018.5.29）
・GRADE system に準拠して大幅改定
6.0→6.1（2018.6.29）
・表現の修正

Critically appraised topic for Systematic review

Reviewer: 年 月 日

authors：

title：

citation：

PubMed PMID：

1. 論文の PICO は何か？

P：

I：

C：

O：

2. コクランレビューか？

□コクランレビューである □コクランレビューでない

3. GRADE approach を用いているか？

□GRADE approach を用いて作られている □GRADE approach は用いられていない

4. 全ての研究を網羅的に集めようと努力したか？

①データベースは？ □MEDLINE □EMBASE □CENTRAL □Cochrane Library □CINAHL □ISI Web of Science □Google scholar その他（＿＿＿＿）

②検索語（＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿）

③期間（＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿）

④どのような種類の研究を調べたか？ □RCT/quasi-RCT/non-RCT/CCT □SR/MA/CDSR □コホート研究 □症例対照研究 □診断研究 □その他の種類の研究（＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿）

⑤参考文献まで調べたか？ □参考文献まで調べた □参考文献は調べなかった □不明

⑥個々の研究者や専門家に連絡を取ったか？ □連絡を取った □連絡を取らなかった □不明

⑦出版されていない研究も探したか？ □探した □探さなかった □不明

⑧英語以外で書かれた研究も探したか？ □探した □探さなかった □不明

5. 全ての研究が網羅的に集められたか？

□研究数が 9 件以下である →明らかな出版バイアスがあるとは言えない

□研究数が 10 件以上である

○ファンネルプロットを用いて出版バイアスの有無を検討している

□ファンネルプロットは対称→出版バイアスはない □ファンネルプロットは非対称→出版バイアスがある

○ファンネルプロットは用いられていない

□出版バイアスは明らか □出版バイアスはなさそう □不明

6. 集められた研究の risk of bias は評価されたか？

①複数の評価車によって評価されたか？

□複数の評価者→＿＿人 □単独の評価者 □不明

→評価者間で評価のくい違いが生じた場合

○合意を形成して最終的に評価を下している ○合意を形成せず，各評価者の判断を個別に記載している

○その他（＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿）

②どのような評価基準で評価されたか？

□Cochrane risk of bias tool で評価した

□Jadad score で評価した

□それ以外の評価基準で評価した →どのような評価か？（＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿）

介入研究の場合の評価するべき項目

○ランダム割付け順番の生成
○割付け方法の隠蔽化
○研究参加者と治療提供者のマスキング
○アウトカム評価者のマスキング
○不完全アウトカムデータ
○選択されたアウトカムの報告
○その他のバイアス

□明確な基準はない

7. 結果の評価（PICO 毎に評価）

①採用された研究の種類（＿＿＿＿），数（＿＿）件，症例数（＿＿＿）人

②統合された結果に有意差があるか？

③統合された結果の大きさは（点推定値と信頼区間）？

④集められた研究に異質性 heterogeneity はあるか？○Cochran Q（カイ二乗検定） ○I^2 統計量（＿＿）%

⑤異質性が高い場合には，高い原因は何か？

⑥異質性が高いものに対してサブ解析，感度分析が行われていれば，その結果を評価する