Beginners' Training Sheet for Clinical Trial

ver.6.2 by last updated on February 7, 2013
南郷　栄秀　Eishu NANGO, MD, PhD
The SPELL　http://spell.umin.jp

このシートは初めて臨床試験の論文を読むためのものです．臨床試験の定義と論文の構造にも触れながら，論文を読む上でのポイントを解説しました．

なお，このシートに関する質問，改善点などは，制作者まで直接お願いします．また，制作者は著作権を保持し，無断転載を禁止します．再配布に制限はしないつもりですが，再配布する際は制作者までご一報ください．

0 治療法・予防法の効果を検証するための研究デザインとは？

0-1）臨床試験（介入試験）とは？

臨床試験 trial には幾つかの種類の研究デザインがあるが，いずれも，介入（ある治療法や予防法）の**治療効果，予防効果**や**比較的頻度の高い害**を調べるために用いられる．

定義：患者に対してある種の介入を加えてその効果・害をみるタイプの研究デザイン．コホート研究や症例対照研究などの観察研究に対して，介入研究とも呼ぶ．また，時間経過中に観察点が複数あり，縦断研究でもある．論文には **RCT(randomized controlled trial)**または **prospective clinical trial** と書かれていることが多い．

目的：①**治療効果，予防効果**を調べる
②**害**を調べる

分類：臨床試験にはいくつかの研究デザインがある．
One arm trial：対象患者に何らかの介入を加えて効果をみるデザイン．対照群はおかない．
Historical controlled trial：過去の研究データから発生率などのデータを対照群として引用する．
non-Randomized controlled trial：患者を介入群と対照群に（ランダム割付け以外の方法で）分けてその効果をみる．
Self controlled trial：自分自身をコントロールとしてデータを複数回とりつつ，介入を加える．crossover trial など．
Randomized controlled trial; RCT：介入群と対照群をランダムに割り付けて比較する．

治療効果・予防効果，比較的頻度の高い害以外の，診断や予後・原因，頻度などのカテゴリーを扱う場合には，臨床試験が最も適切な研究デザインとはいえない．

他のカテゴリーの場合：
診断→横断研究
予後・原因→コホート(Cohort)研究
害→症例対照研究（稀な場合），RCT（比較的頻度が多い場合）
頻度→横断研究

0-2）臨床試験に含まれる研究デザインのいろいろ～RCT とは？

例えば，「高血圧患者が降圧剤を飲むと，脳卒中が減るだろうか？」という命題を検証する場合を考えると，PICO（p.3 参照）は以下の通りになる．

P：高血圧患者が
I：降圧剤を服用するのは
C：降圧剤を服用しないのと比べて
O：脳卒中が減るか？

これを検証するためには，以下のような研究デザインが考えられる．

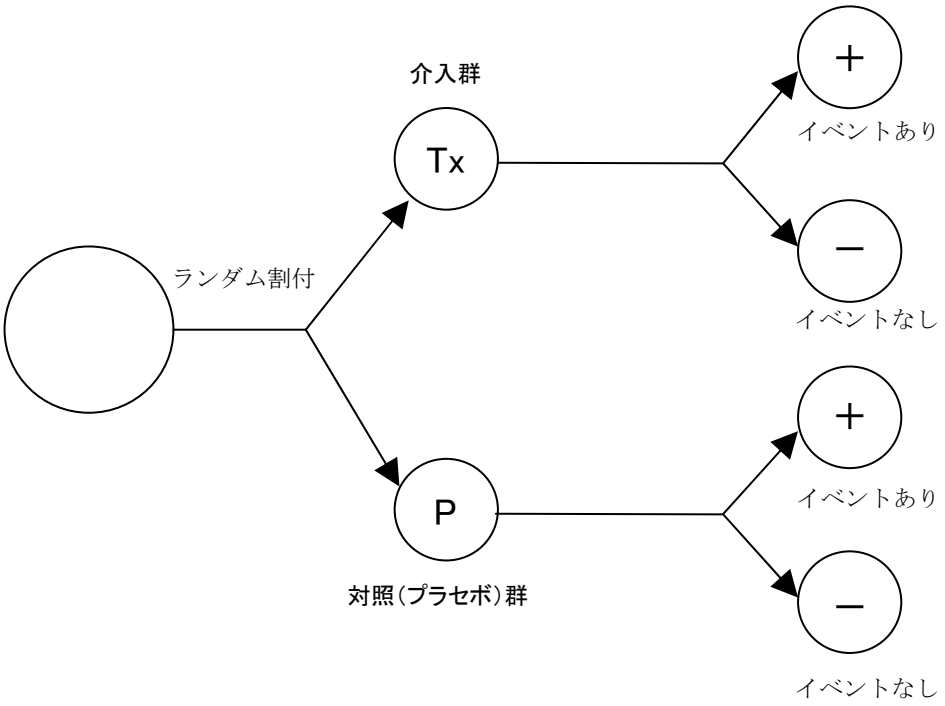
①降圧剤を一定期間服用して脳卒中にならなかった 1 人の高血圧患者がいた
→これでは，その患者が脳卒中にならなかったことが偶然だった可能性が否定できない．このような研究の報告を「**症例報告 Case report**」と呼ぶ．

②降圧剤を一定期間服用して脳卒中にならなかった何人かの高血圧患者がいた
→これでは，個々の患者についての効果から，どのような傾向があるかを推測することしかできない．また，効果のあった患者のみを報告している可能性もある．このような研究を「**症例集積研究 Case series study**」と呼ぶ．

③複数の高血圧患者を集めて降圧剤を一定期間投与したところ，その結果，誰も脳卒中にならなかった
→これでようやく，降圧剤が，特定の集団（この場合は高血圧患者）に対してどのくらい効果があるかが分かる．しかしまだ脳卒中の予防が本当に降圧剤によるものとは断言できない．同じ時期に行っていた別の治療が功を奏した可能性もあれば，あるいは，もともと何も治療しなくても自然に脳卒中が予防できたのかも知れない．こうした研究を「**比較対象をおかない臨床試験 Single arm clinical trial**」と呼ぶ．

④複数の高血圧患者を任意に 2 つの群に分け，一方には降圧剤を一定期間投与して，もう一方には降圧剤を投与せずに一定期間経過を追った結果，降圧剤を投与した群の方が，脳卒中の発生率が低かった
→このように降圧剤を投与しない群と比較することで，降圧剤により脳卒中の発生率が減少した可能性が高まる．しかし，これでもなお問題が残る．初めに患者を 2 群に分けた際に，実は，降圧剤を投与した群の患者の方が，脳卒中になりにくかったのかも知れないし，研究者が，降圧剤に有利な結果を出すために，健康に注意を払うことができそうな患者を，降圧剤を投与する群に意図的に割り付けた可能性もある．このような研究を「**非ランダム化比較試験 non-Randomized controlled trial**」と呼ぶ．

⑤複数の高血圧患者をランダムに 2 つの群に分け，一方には降圧剤を一定期間投与し，もう一方には降圧剤を投与せずに一定期間経過を追った結果，降圧剤を投与した群の方が脳卒中の発生率が低かった．
→④の問題点を解決するには，降圧剤を服用するか否かという点を除き，両群の条件（＝背景因子）は全く同等にする必要がある．2 つの群の背景因子を同等にするための唯一の方法は，ランダム割り付けと呼ばれるもので，一人一人の患者をどちらの群に割り付けるかを，サイコロを振る要領でランダムに決定する．この方法で行われた研究を「**ランダム化比較試験 Randomized controlled trial（以下，RCT）**」と呼ぶ．



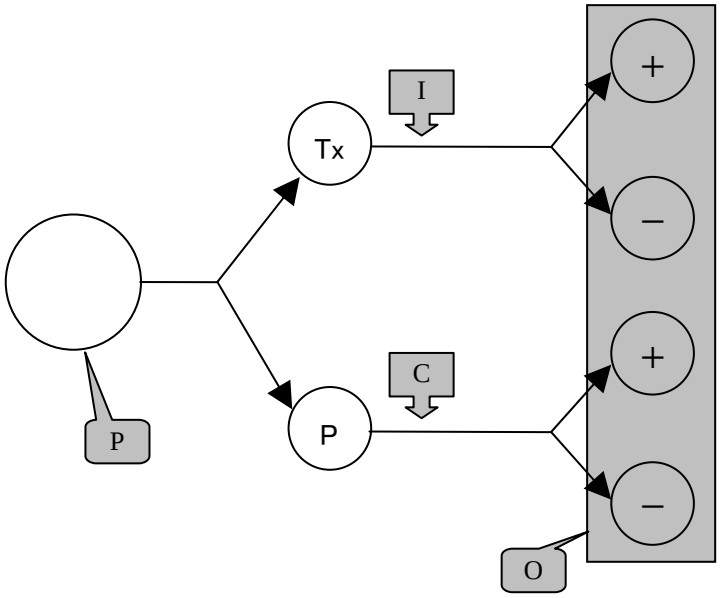
RCT 論文の構造

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や研究結果は，「要約 abstract，summary」に記載されているが，この部分だけでは情報が不十分なことも多く，論文の「方法 methods」の項で詳細を確認することが必要である．

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．



1-1）論文の patient

患者 P(Patient)：患者選択・エントリー（参加）の基準は何か？

組み入れ基準 inclusion criteria：

除外基準 exclusion criteria：

記載がある可能性の高い場所
「要約 Abstract」の「方法 methods」
本文の「方法 methods」にある「患者 patients，参加者 participants」のはじめの方
記載を見つけるためのキーワード
patients／participants
include／inclusion／exclude／exclusion
eligible／enrolled
＜例＞
Patients were **eligible** if they had had an acute myocardial infarction or had a hospital discharge diagnosis of unstable angina between 3 and 36 months before study entry.

通常，介入試験は，何かの疾患に罹患している人や何かの状態（健常者を含む）に対して行われる．従ってその疾患や状態が論文中に明確に定義されているはずである．

通常，試験へのエントリー（参加）は，inclusion criteria にそって大まかに対象者を集め，その中から exclusion criteria に基づいて不適合症例を除外する，という 2 段階で行う．

多くの場合，「方法 methods」の「患者 patients，参加者 participants」の項に記載されている．

1-2）論文の intervention，comparison

介入 I(Intervention)：

比較 C(Comparison)：

記載のある可能性の高い場所
「要約 Abstract」の「方法 methods」
本文の「方法 methods」にある「介入 intervention，治療 treatment，研究デザイン study design」の，患者についての記載の後
記載箇所を見つけるためのキーワード
assign／receive
（「random」の直後にあることが多い）
＜例＞
After stratification according to the qualifying event (myocardial infarction or unstable angina) and clinical center, patients were randomly **assigned** to receive either 40 mg of pravastatin (Pravachol, Bristol-Myers Squibb) or matching placebo once daily.

次に，研究対象としている患者や参加者に対して，どのような介入（治療）をしているかを調べる．これは，臨床試験では通常，

介入 I(Intervention)：その研究で効果を評価したい治療法
比較 C(Comparison)：placebo または他の治療法

となる．これ以外の介入は両群で同等でなければならない．同等でないと，それが原因で結果が左右される（バイアスとなる）可能性がある．従って，介入以外の治療が同等かどうかも，チェックする．

1-3）論文の outcome

結果 O(Outcome)：

記載がある可能性の高い場所
要約「Abstract」の「方法 methods」
本文の「方法 methods」にある「結果測定 Outcome measurement，研究デザイン study design」
記載箇所を見つけるためのキーワード
main outcome／primary outcome
primary endpoint
＜例＞
The primary study outcome was death from CHD. Deaths from CHD were further classified as death due to fatal myocardial infarction, sudden death, death in the hospital after possible myocardial infarction, or death due to heart failure or another coronary cause.

論文によっては，治療効果を評価するものとして，複数の効果指標を outcome として設定している．このうち，最も重要なものは主要エンドポイント（primary endpoint，main outcome）と呼ばれる．研究はこの主要エンドポイントを評価するために計画されている．多くの場合，「方法 methods」の「結果 outcome，評価 assessment」の項に記載されている．

ここでチェックする「結果 outcome」とは，あらかじめ研究開始前に決められた，その研究で測定しようとしている効果の“指標”のことである．実際の効果の大きさについては，「8．結果の評価」で評価する．したがって，まだ「結果 Result」の項を見てはいけない．

2 ランダム割付けされているか？

□ランダム割付け randomized
□非ランダム割付け non-randomized
割付けの方法：

ランダム割付けが隠蔽化 concealment されているか
□隠蔽化されている
□隠蔽化されていない
□不明

記載がある可能性の高い場所
タイトル
「要約 Abstract」の「方法 methods」
本文の「方法 methods」にある「介入 Intervention，研究デザイン study design」の，介入方法の説明の前後
記載箇所を見つけるためのキーワード
random／randomly／randomize／randomization
conceal／concealment
independent center／sequentially numbered／identical appearance／opaque／sealed envelopes
＜例＞
After stratification according to the qualifying event (myocardial infarction or unstable angina) and clinical center, patients were **randomly** assigned to receive either 40 mg of pravastatin (Pravachol, Bristol-Myers Squibb) or matching placebo once daily.

ランダム割付け random allocation というのは，研究にエントリー（参加）した患者を偏りなく複数の群に分ける作業である．ここで，「ランダム抽出」と「ランダム割付け」を混同してはいけない．「ランダム抽出」は論文の外的妥当性を保証するもの，「ランダム割付け」は交絡因子を排除することで内的妥当性を保証するものである．RCT で最も重要なのは本当に「ランダム割付け」されているかどうかである．

ランダム割付けの方法には，中央割付方式，封筒法などがあるが，中央割付方式が最も優れている．曜日による割付け，患者番号による割付け，患者の誕生日による割付けなどは，一見ランダムに割付けられているようであるが，恣意的なものとなりうる．従って，これらは準ランダム割付け quasi-random allocation といって厳密にはランダム割付けには含まれない．

多くの場合，タイトルや「方法 methods」の「介入 intervention，研究デザイン study design」の項に記載されている．Random という文字を見つければ，ランダム割付けされている可能性が高いが，厳密に random allocation されているか，quasi-random allocation なのかを評価する必要がある．

2-1）ランダム割付けの隠蔽化

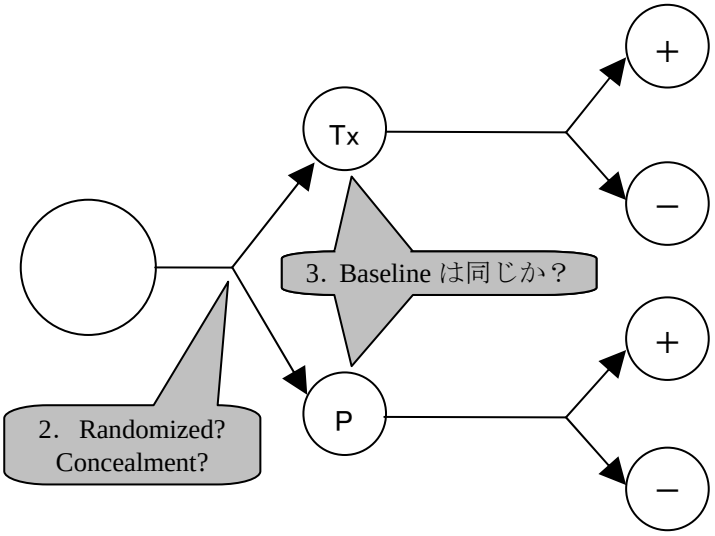
ランダム割付けを行う際に注意しなければならないのは，ランダム割付けが**隠蔽化**（いんぺい） concealment されているかどうかである．臨床試験において患者を研究に組み入れて介入に割付けようとしているときに，それまでに別の患者に割り付けた介入が何だったのかを知っていると，意識的にあるいは無意識的に患者を組み入れるか否かを選んでしまい（＝選択バイアス selection bias），その結果両群のバランスを崩してしまうことがある．これを防ぐため，既に臨床試験に組み入れた患者がどの介入に割付けられたかを，患者を研究に組み入れ，割付けをする者に知られないようにすることが望ましい．

隠蔽化 concealment とマスキング masking とはよく混同される．隠蔽化は，研究開始前に患者を研究に組み入れる者が，これから組み入れようとする患者がどちらの群に割付けられるか予想できないようにすることであり，一方，マスキングは，研究開始後に患者や医師，結果の評価者などが割付け内容を知らないことである．つまり，隠蔽化とマスキングは互いに独立したものであり，隠蔽化されているのにマスキングされていない，または隠蔽化はなされていないがマスキングはなされているという研究も存在する．隠蔽化とマスキングの違いのポイントは，隠す時点が違うということと，隠されている人が違うという 2 点である．Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 では，中央割付方式（電話，web ベース，薬局でランダム割付けをする），外見の分からない（identical appearance）連続番号の薬コンテナ，封のされた不透明な連続番号の封筒のいずれか，またはそれと同等の方法でランダム割付けが行われた場合は隠蔽化されていると判断して良いとされている．

2-2）層別化 stratification

ランダム割付けを行った場合でも，特に症例数が小さい場合は，偶然に両群の間に差が生じることがある．例えば，両群間で疾患の重症度に重大な違いが生じてしまった場合など，結果の解釈が難しくなる．これを解決するために，**層別化** stratification した上でランダム割付けするという方法がある．割付け前にあらかじめ，性別や疾患の進行度など，両群間で違いを生じさせたくない因子ごと（層）に患者を分けておき，それぞれの層の中でランダム割付けを実施して組み合わせるというものである．これにより，群間には，層別化した因子についてはバラツキがなくなる．

但し，層別化は最小限にすべきである．層別化する因子が増える毎にそれぞれの層の症例数が少なくなってしまい，その中でランダム割付けを行うことが困難になるからである．



3 Baseline は同等か？

□群間に差がない
□群間に差がある
差がある場合はどこにあるか？

結果に影響を与える可能性のある因子は全て検討されているか？
□検討されている
□不足しているものがある

記載がある可能性の高い場所
本文の「結果 Results」の最初の段落
Baseline の表（多くの場合は「Table1」）
記載箇所を見つけるためのキーワード
baseline
characteristic／factor／feature
similar／not significant／no difference／well balanced／evenly distributed
＜例＞
The two groups were very **well balanced** in terms of base-line characteristics (Table 1).

ランダム割付けされているので，群間に差がないように割付けられているはずだが，本当に差がないかの確認をする．Baseline の比較がなされている表（通常 table1）をみて，群間に差がないかどうかチェックする．ランダム割り付けをした結果として差がないかを評価するので，本文中では，「結果 Results」の冒頭に記載されていることが多い．Baseline characteristics の項を見て，**similar**，**not significant**，**well balanced** 等と書いてあれば，差がないといえる．このとき，交絡因子（結果に影響を与えると考えられる因子）になりうるもの（介入以外の治療も含めて）がすべて表に掲載されているか検討する．ここでは，研究で扱われている疾患に対する医学的知識が必要である．

4 全ての患者の転帰が outcome に反映されているか？

4-1）ITT 解析か？

☐Intention-to-treat 解析されている
☐Intention-to-treat 解析されていない
→されていない場合，それは結果をくつがえしうるほど重大か？
〇結果をくつがえしうるほど重大である
〇結果をくつがえしうるほど重大ではない

記載がある可能性の高い場所
「要約 Abstract」の「方法 methods」
本文の「方法 methods」にある「統計学的解析 Statistical analysis」の前半
記載箇所を見つけるためのキーワード
Intention-to-treat／ITT
＜例＞
All analyses were performed on an **intention-to-treat** basis.

ITT とは intention-to-treat の略である．和訳すると，「意図された通りの治療」となる．つまり，**研究の初めに治療を割付けた通りに解析を行うことが ITT** である．多くの場合，「方法 methods」の「統計学的解析 statistical analysis」の項にズバリ intention-to-treat と記載されている．記載がない場合は，baseline の表（割付時，通常 Table 1）と結果の表（通常 Table 2）の n（症例数）を比較する．

ITT 解析の原理原理は以下の 3 つである．

1. 一度特定の群に割りつけたら，実際の治療が異なっても，割りつけた群のままで解析を行う．
2. 全患者でアウトカムを測定する．
3. ランダム割付けを行った全患者を解析に含める．

ITT の目的は，「ランダム割付け」の保持である．治療群の患者が実際には治療を受けなかったり，逆に placebo 群の患者が治療を受けなかったりした場合，そこには必ず理由があるはずである．たとえば，治療薬の副作用により治療続行が不可能と判断されれば，途中で投薬が中止される場合もある．一方，placebo 群の患者が Masking をされていたとしても自分は placebo を与えられていると信じ込み，他の医療機関を受診するなどして実薬を服用し始めるかも知れない．こうした場合でも，初めの割付け通りに解析を行うことで，治療の効果が真の効果よりも過小評価されると考えられるため，得られた結論がより頑強となる．すなわち，ITT 解析で効果があると評価された治療法は，より不利な条件でも効果があると考えられ，真には効果がないという可能性が低くなるのである．

なお，現在まだ ITT の概念自体が流動的である．ITT に関連した概念には次のようなものがある．

①ITT（狭義の ITT）：厳密な意味での ITT で，理由に関わらず全てを含めて解析する．ただし，これは現実的には，困難である．そのために，実際には，次に説明する FAS で解析されることが多い．

②FAS (Full analysis set)：一度も介入を受けていない患者のみ除外する．ITT があまりにも厳密すぎて，これに従って臨床研究を実行することが困難なため，1 つの妥協策として提唱された概念．但し，除外する理由を割り付け開示前にプロトコールに明記することが条件となる．そうでなければ，特定の群へ割り付けられたために治療を遵守しないという症例が発生し，それを除外することはバイアスとなる可能性があるからである．

③Per protocol analysis：プロトコールに従った患者だけ選んで解析する．

④GCP 不適合例：FDA の指針などに基づき，倫理的に問題のある症例・施設を除外して解析する．

通常，ITT と呼ばれるのは①と②である．本来は①のみを ITT とするべきだが，②ではいけないという結論にはなっていない．これら広義の ITT で得られた結果と比較するために，③を行ってみるとよい（論文中に解析結果が示されていることもある）．ただし，副作用の論文では①が用いられるべきである．④の扱いについてはまだ結論が得られていない．

baseline の表と結果の表で n が一致しない場合は本文から脱落理由を読みとる．合理的な脱落であれば（評価すべき outcome が起こり得ない状態にあった場合など），n が一致しなくても ITT であるとするのが Per protocol analysis である．

4-2）結果に影響を及ぼすほどの脱落があるか？

□ない
□ある
追跡率＝結果の症例数／割付け時の症例数＝

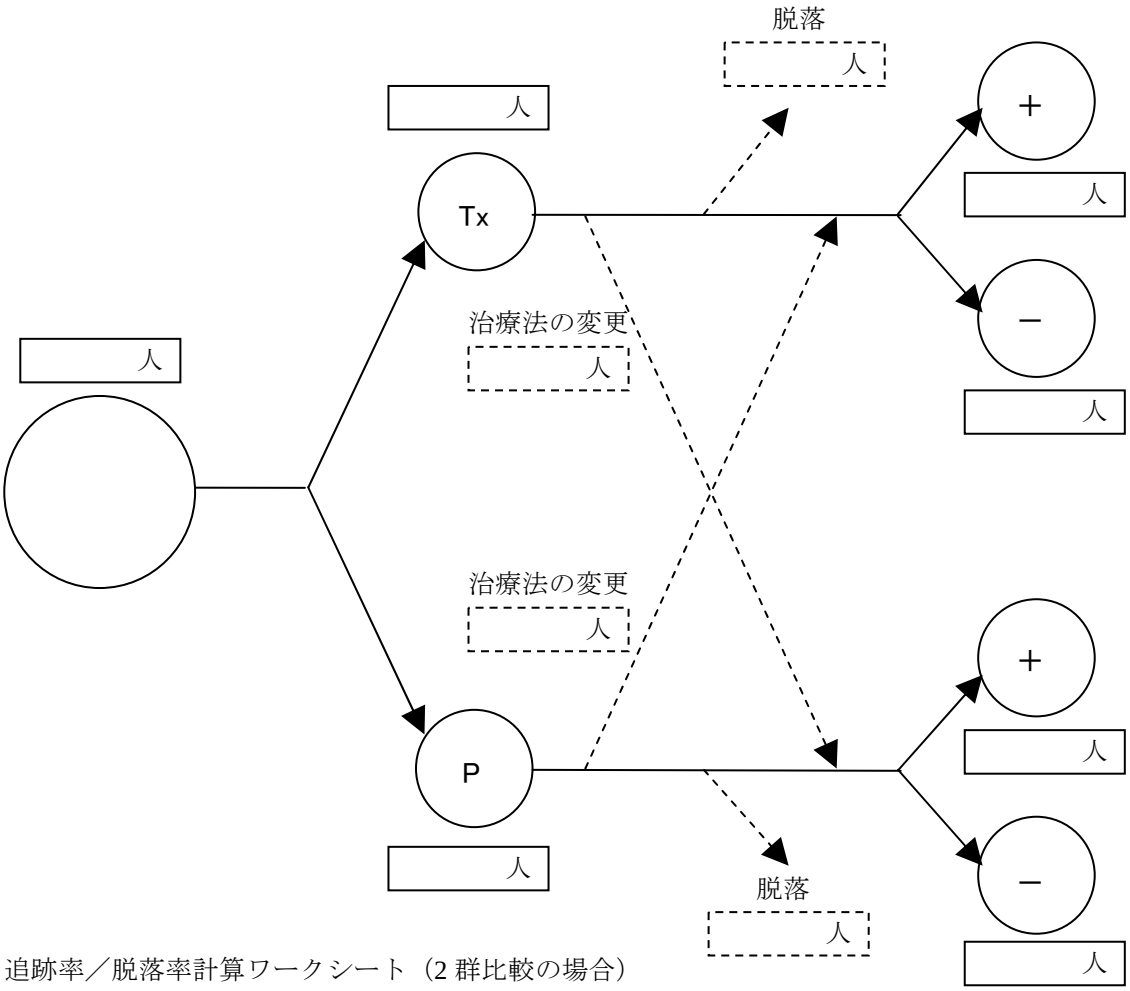
□不明

記載がある可能性の高い場所
本文の「結果 Results」の最初の段落
記載箇所を見つけるためのキーワード
follow up rate, lost to follow up
withdraw／stop／compliance
＜例＞
After one year, after three years, and at the end of the study, 6 percent, 11 percent, and 19 percent, respectively, of the patients randomly assigned to treatment with pravastatin had permanently **stopped** taking the study drug, whereas 3 percent, 9 percent, and 24 percent of those assigned to placebo had begun open-label therapy with a cholesterol-lowering drug.

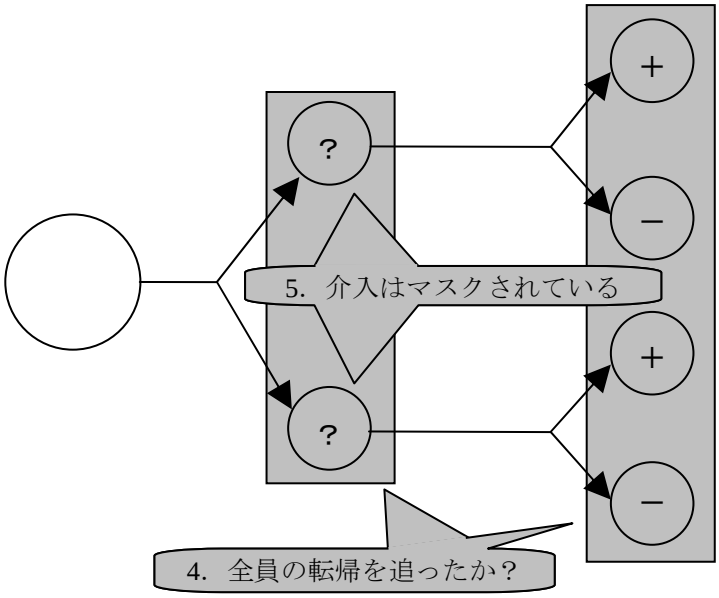
脱落とは，最終的なアウトカムが起こったか否かが不明となった症例を指す．ITT 解析されていたとしても，脱落が多く追跡率が低ければ，ランダム割付けにより均等に割付けたはずの各群に偏りが生じてしまう．特定の群だけ，脱落が多くなる可能性があるからである．脱落した症例の数が，研究結果に影響を及ぼすほど大きくなってしまうと問題である．

割付け時の症例数のうち，結果が判明している症例数の割合が追跡率である．追跡率が 80%に達しないときには，ランダム割付けが保持されているとは言えず，内的妥当性が疑わしくなる（ただし，この 80%という数字は絶対的なものではない）．多くの場合，「結果 results」の冒頭に記載されている．

なお，脱落率＝1－追跡率である．



追跡率／脱落率計算ワークシート（2 群比較の場合）



5 マスキング（盲検化）されているか？

マスキング（盲検化）されているのは誰か？
- **〇患者，参加者**
- **〇介入（治療）実施者**
- **〇Outcome 評価者（結果を判定する者）**
- **〇データ解析者（統計処理を行う者）**

□四重盲検 Quadruple masking
□三重盲検 Triple masking
□二重盲検 Double masking
□一重盲検 Single masking
□マスキング（盲検化）なし Open labeling
□マスキング（盲検化）不可能
□不明

記載がある可能性の高い場所
タイトル
「要約 Abstract」の「方法 methods」
本文の「方法 methods」にある「介入 Intervention，研究デザイン study design」の，介入方法の説明の前後
本文の「緒言 Introduction」の最後
記載箇所を見つけるためのキーワード
mask／masking／blind／blinding／aware
open／open label／unaware
＜例＞
In a **double-blind**, randomized trial, we compared the effects of pravastatin (40 mg daily) with those of a placebo over a mean follow-up period of 6.1 years in 9014 patients who were 31 to 75 years of age.

一人一人の患者がどのような介入を受けているのか，患者や臨床医は知らないことが望ましい．また，評価者が介入を認識していると，outcome の評価を無意識のうちに歪めてしまうことがある．しかし，現実的にマスキングできない場合もある．

理想的な臨床試験では，患者（参加者）・介入実施者・outcome 評価者・データ解析者が一様にマスキング（盲検化）される．一重盲検・二重盲検・三重盲検・四重盲検という用語はこれらのマスキングの組み合わせの数を表現したものである．同じ二重盲検といっても，研究によって誰に対して行われているかが異なるため，マスキング（盲検化）が行われたのが誰なのかを確認すること重要である．

多くの場合，タイトルや「方法 methods」にある「介入 intervention，研究デザイン study design」の項に記載されている．最も詳しく書いてあるのは，「方法 methods」の「介入 intervention，研究デザイン study design」の項であるが，その他の場所に書かれている場合は，通常マスキングが何重かだけが書かれており，その対象が誰かの記載はないことがほとんどである．しばしば「緒言 Introduction」が唯一のマスキングについての記載場所となる．また，マスキングが不可能であることが明らかな場合は，論文中にマスキングに関する記載がないこともある．

他に，研究結果を左右する可能性のある効果として，プラセボ効果とホーソン効果が知られている．

5-1）プラセボ効果 placebo effect

患者は治療を受けていると思っている方が，思っていない場合よりも状態が良くなる．このプラセボ効果は臨床試験の潜在的なバイアスとして広く知られている．従って薬物療法において，このバイアスをなくすため，偽薬 placebo を用いて，真の薬剤の効果以外の影響が同等になるよう工夫されている．

5-2）ホーソン効果 Hawthorn effect

人間は仕事をする際，誰にも見られていないとさぼったり手を抜いたりするが，誰かに見られているとそういったことは控える性質がある．観察されているという意識は，作業者に緊張感を与え，作業により集中させる．もともと，このように実際の効果よりも高めるものをホーソン効果と呼ぶが，実は臨床試験では逆の効果を生じる場合もある．つまり，臨床試験に参加しているのできっとよくなるであろうという意識が働き，生活スタイルが乱れてしまった場合など，効果は実際より低下してしまう．

5-3）PROBE 法

PROBE 法とは，**Prospective, Randomized, Open-labeled Blinded Endpoints**（前向きランダム化オープンラベル（非盲検）試験）の略．患者，参加者と介入実施者（医師）の二者は，経過中，治療群と対照群のどちらに割り付けられたかを知っているが，outcome 評価者はそれを知らずに outcome を評価する方法である．PEOBE 法では，介入実施者が割り付け内容を知っているため，恣意的に脱落させる危険性があり，解釈する際には特に脱落の評価に注意が必要である．特に，outcome の評価に評価者の主観が入り込むようなソフトエンドポイント（脳卒中発症率，血行再建術施行率，喘息発作による入院率など）が設定されている場合の PROBE 法では，outcome の評価頻度を調節することにより outcome 発生率を変えることができてしまう．

6 症例数は十分か？

□結果に有意差がある　→症例数は十分
**　〇サンプルサイズは計算されている**
**　〇サンプルサイズは計算されていない**
□結果に有意差がない　→症例数は十分かどうか不明
**　〇サンプルサイズは計算されている**
**　　研究に参加した人数は計算されたサンプルサイズを超えているか？**
**　　〇超えている　　→症例数は十分**
**　　〇超えていない　→症例数は不十分（症例数増加で有意差が出る可能性あり）**
**　〇サンプルサイズは計算されていない**
**　　症例数（各群：　　合計：　　）**
**　　イベント発生率：　　%**
**　　効果：　　%**
　　α：
　　β：
□不明

記載がある可能性の高い場所
本文の「方法 methods」にある「統計学的解析 Statistical analysis」の後半（稀に前半）
記載箇所を見つけるためのキーワード
sample size
calculate
α／alfa／alfa level／p value
power
＜例＞
The study was designed to have 80 percent **power** to detect a reduction of 18.3 percent in the risk of death due to CHD at five years, with a two-sided **P value** of <0.05. The trial was planned to continue until 700 deaths from CHD had occurred unless it was stopped early.

サンプルサイズとは，研究に必要な症例数を意味する．通常，臨床試験ではあらかじめ，治療効果の差を検出するために必要な症例数をサンプルサイズとして計算しておくものである．たとえば，症例数が少なすぎると，本当は治療効果があるにも関わらず，統計学的には差がないとみなされてしまう．従って，結果に差があったときは症例数は十分と言えるが，結果に差がなかったときには，サンプルサイズが少なすぎたことが原因かどうかを検討する必要が出てくる．サンプルサイズの計算は，多くの場合，「方法 methods」の「統計学的解析 statistical analysis」の項に記載されているが，記載がないこともある．その場合は，研究デザインに関する別の論文が研究開始時に既に発表されていることがある．

サンプルサイズの計算には，以下の 4 つの項目を設定しておく必要がある．

①（対照群の）イベント発生率
②介入をした場合に期待される効果
③α level：α error を起こす確率
④β level：β error を起こす確率，power＝1-βで示されることもある

イベント発生率と期待される効果は過去の研究などをもとに設定される．

α error とは「本当は差がないのに，差があると勘違いしてしまう誤り」である．α level は通常 0.05 に設定されることが多く，研究結果において p 値がこれを下回ると，α error の起こる確率が低く，有意差ありと判断される．α level が変われば，当然 p 値の判断基準も変わる（α level を 0.01 と設定した場合，有意の基準も p＜0.01 となる）．「あわてん坊のα error」と覚える．

一方，β error とは，「本当は差があるのに，差がないと勘違いしてしまう誤り」である．例えば，β level が 0.2（＝power が 80%）である研究において，研究結果が有意差なしとなった場合，20%の確率でβ error が起こっている可能性がある，ということである．つまり，本当は差があるのに，20%の確率で差がないという結果が出てしまう．裏を返せば，実際の結果に有意差があれば，少なくともβ error はないと言える．「ぼんやり者のβ error」と覚えると良い．

7 結果の評価

臨床試験の outcome に用いられる指標は，次の 2 種類に分けられる．

1）時間軸に垂直の指標：発生率，発症率，生存率，死亡率，治癒率など．

2）時間軸に平行な指標：発症までの期間，生存期間，治療期間など．

1 つの臨床試験において，同じ outcome を時間軸に垂直な指標と平行な指標の両方で示すこともある．

1）時間軸に垂直な指標

介入が始まってから一定期間後のある時点における outcome の発生率が指標となっている場合，関連する以下のような指標が計算できる．

追跡期間＝

介入群の発生率＝

対照群の発生率＝

RRR＝

NNT＝

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

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

介入群イベント発生率 experimental event rate ： EER＝a／(a＋b)

対照群イベント発生率 control event rate ： CER＝c／(c＋d)

相対リスク relative risk ： RR＝EER／CER

相対リスク減少率 relative risk reduction ： RRR＝(CER－EER)／CER

絶対リスク減少率 absolute risk reduction ： ARR＝CER－EER

治療必要数 number needed to treat ： NNT＝1／ARR

1)追跡期間

臨床試験では，outcome が生じるまでに十分な時間を必要とする．追跡期間が十分でないと outcome が生じるに至らず，結果の解釈を誤る可能性がある．また，同じ治療法であっても，outcome を測定する時間によってその発生率は異なる．従って，結果の解釈には，追跡期間がどれくらいであったかが重要である．多くの場合，「結果 results」の冒頭に記載されている．

2)相対評価と絶対評価

Primary outcome（Primary endpoint）の発生率が記載されていれば（表を見るのが早い），そこから相対評価である RRR と絶対評価である NNT を計算する(RR や RRR は論文中に示されていることも多い)．

RR や RRR は，介入そのものの効果を見た指標である．RR が 1 を下回ると，介入群の方が対照群よりも効果があり，逆に 1 を上回ると介入群の方が効果が劣る（＝害がある）ことになる．なお，相対評価は，発症率や追跡期間に影響を受けることがなく，介入固有の効果の大きさを示している．

これに対し，ARR や NNT は，発症率に左右されるため，臨床現場でより重要となる指標である．**NNT は 1 人の望ましくない outcome を防ぐのに介入が必要となる人数**を示す．したがって，小数点以下を切り上げて整数で示す．NNT は大概的には 2 桁で有用性があり，1 桁であればかなり効果が期待できるとされる．逆に介入することでかえって効果がマイナスに現れたり，害を outcome としている場合は，number needed to harm; NNH と呼ぶ．NNT は追跡期間によって変化するので，必ず追跡期間を併記する．

重要なことは，介入群の発症率，対照群の発症率，相対評価での効果の大きさ，絶対評価での効果の大きさの 4 つの数値を全て見て総合評価することである．有意差の有無と，臨床上の効果は別物である．特に重要なのは RRR や NNT などの推定値の大きさであり，推定値が小さければ，有意差があったとしても効果は小さいということになり，実際の医療行動を変えるだけのインパクトがあるとは言い難い．

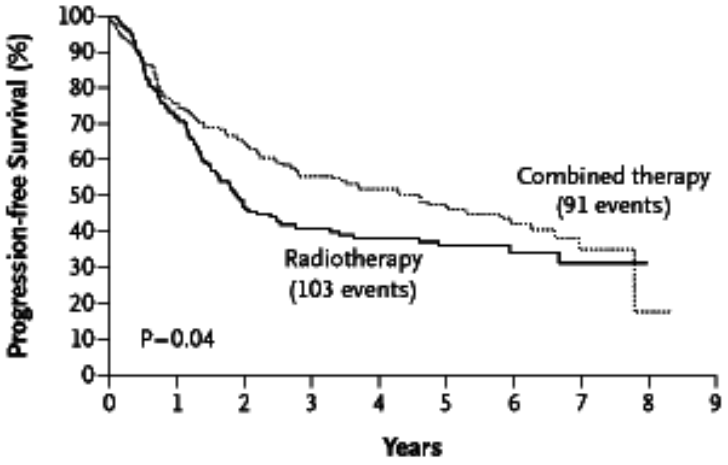
なお，時間軸に垂直な指標でも，“率”が求められない場合（血圧や検査値などの測定値，症状スコアなど）は，RRR や NNT は計算できない．

3)95%信頼区間 confidence interval と p 値 p-value

95%信頼区間 95%CI や p 値 p-value の記載があれば，それらも評価する．95%信頼区間とは，研究結果の真の値が 95%の確率で存在する範囲を指す．95%信頼区間は，RR や RRR では 1 を，ARR では 0 を，NNT では∞を含まないときに「有意差がある」という．95%信頼区間は最も過小評価した場合と，最も過大評価した場合の効果についても考える．

一方 p 値は介入群と対照群の効果が同等だと仮定したときに，偶然効果に違いがあるという結果が出る確率である．例えば p=0.02 なら介入群と対照群の効果に偶然違いが出る確率は 2%である（その確率は稀なので両群の効果の違いは必然といえる）．p 値が 0.05 よりも小さい場合に，「有意差がある」という．

2）時間軸に平行な指標



No. at Risk									
Radiotherapy	167	119	73	57	45	30	18	9	0
Combined therapy	167	125	105	85	66	42	29	10	1

図 1：Kaplan-Meier 曲線の例（NEJM2004;350:1945 から）

結果の指標として，outcome が発生するまでの時間（生存期間や治療期間など）の平均値や中央値が用いられている場合もあるが，これらは時間軸に平行な指標となる．これら時間データを対象とする解析手法を総称して，**生存分析(survival analysis)**という．この方法は，outcome の発生が 1 回しかない場合か，複数回ある場合は始めの 1 回だけをカウントする場合に用いることができる．

RCT の場合，通常，生存期間や治療期間の始点はランダム割付け時となる．1）（時間軸に垂直な指標）との大きな違いは，患者の転院などにより，研究の途中で出た脱落症例の扱いである．生存分析では，脱落した症例も完全に除外することなく，打ち切り（censor）として扱い，研究に参加していた期間のデータを解析に含めることができる．そのため，途中で脱落した症例のデータを捨てずに済む．

生存分析を行うにあたり，まず，経過中の各時点で，まだ outcome が発生していない人の割合を群毎にプロットした **Kaplan-Meier 曲線**（図 1）を描く．具体的には，横軸に時間をとり，縦軸に outcome の発生率に相当する指標（生存率，発症率，死亡率など）をとる．各時点での症例の母数（outcome が発生する可能性のある症例数 No. at Risk）は時間と共に減少するが，論文によっては，グラフの横軸に沿って母数が書かれていることもある．生存率は，outcome が発生しない間は変化せず，outcome が 1 人発生した時点で減少するので，Kaplan-Meier 曲線は階段状になる．

Kaplan-Meier 曲線が描けたら，介入群と対照群の曲線の形状を評価する．時間経過に従って差が開いていくもの（図 2A），一旦差が広がるが，最終的には同じになるもの（図 2B），途中で交叉して，優劣が逆転するもの（図 2C）などが考えられる．

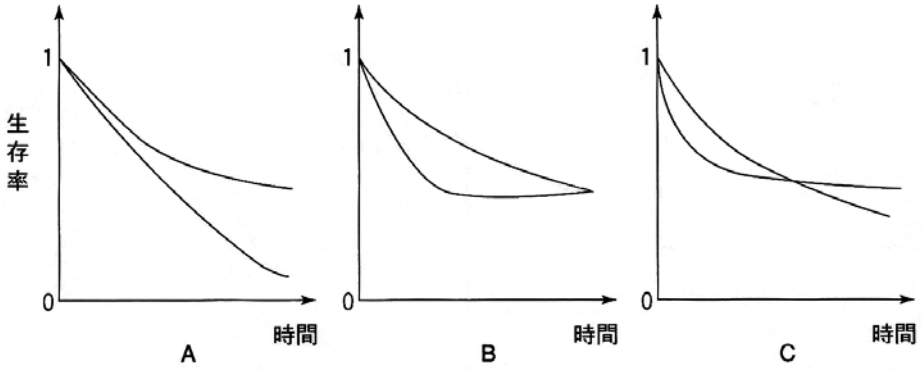


図 2：Kaplan-Meier 曲線の形と有意差の出やすい検定

また，これらの曲線が統計学的に異なるか検定する．この検定により有意差が出た場合，曲線の形状が異なることが示され，治療の効果に差があるといえる．この検定は，各時点での 2 つの曲線の離れ度合いを合計して行うもので，**ログランク検定 Log-Rank test** や**ウィルコクソン検定 Wilcoxon test** が用いられる．ログランク検定は全ての時点の重みを等しいとして合計するため，群間の違いが一定で，時間経過と共に生存曲線の差が広がっていく場合に差が出やすい（図 2A）．一方，ウィルコクソン検定は，母数が研究の早期ほど大きいことを重視して重み付けをしているため，研究の比較的早期に起こる outcome を重視する場合に差が出やすい（図 2B）．途中で交叉してしまうような曲線の場合は，どちらの検定でも差が出にくい（図 2C）．

さらに，Cox 比例ハザードモデル Cox proportional-hazards model を用いて，相対危険度を求める．これは，介入以外に outcome の発生に影響を与える可能性のある因子を排除するための方法である．

参考文献

1）Guyatt GH, Sackett DL, Cook DJ, for the Evidence-Based Medicine Working Group: Users' guides to the Medical Literature. II: How to Use an Article About Therapy or Prevention. A. Are the Results of the Study Valid? JAMA 1993;270(21):2598-2601.

2）Guyatt GH, Sackett DL, Cook DJ, for the Evidence-Based Medicine Working Group: Users' guides to the Medical Literature. II: How to Use an Article About Therapy or Prevention. B. What Were the Results and Will They Help Me in Caring for My Patients? JAMA 1994;271(1):59-63.

3）CASP JAPAN. http://caspjp.umin.ac.jp/materials/RCT21j.pdf.

4）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，11-35.

5）Ttrain ML. http://www.egroups.co.jp/group/Ttrain/.

6）厚生省医薬安全局審査管理課長，臨床試験のための統計的原則 E9，1998 年.

7）Bedenoch D 他著，斉尾武郎監訳，EBM の道具箱，中山書店 2002 年，16-24.

8）浜田知久馬著，学会・論文発表のための統計学，真興交易医書出版部，167-182，1999 年.

9）Higgins JPT, Green S. Cochrane Handbook for Systematic Reviews of Interventions. Version 5.1.0. The Cochrane Collaboration, 2011.

改訂履歴

1.1→1.2LE（2003.3.9）
- 1 時間半～3 時間の初学者向け EBM セッション用のトライアルシートとして要点のみ圧縮

1.2LE→1.3LE（2003.4.1）
- ランダム割付けの隠蔽化についての記載を追加
- 「baseline の比較」はランダム割付けの結果，交絡因子が排除されていることを確認するためのものなので，ランダム割付けの評価後に順番を修正
- “盲検化”という用語よりも“マスキング”を優先とした
- マスキング（盲検化）の評価について，誰がされているかのチェックリストを追加
- 書き込み用のスペースの確保

1.3LE→1.5LE（2003.7.4）
- 書き込み用 CAT sheet を作成

1.5LE→2.0LE（2003.12.11）
- 細かい表現の訂正
- 臨床試験の概念の解説の項において，「治療効果」に加えて，「予防効果」を追加
- inclusion criteria，exclusion criteria に日本語訳を添付
- baseline の比較に全ての因子が検討されているか否かの項を追加
- 追跡率との混乱を避けるため，追跡期間の項を結果の項に移動
- マスキングの項に「マスキング不可能」の欄を追加
- サンプルサイズの項の設問と解答形式を変更

2.0LE→3.0（2004.3.11）
- 「隠蔽化」，「層別化」の記載を追加
- 「ITT」の記載を追加
- 「プラセボ効果」，「ホーソン効果」の記載を追加
- 「結果の評価」の記載を変更
- 書き込み用 CAT sheet の改変（1 ページ化）

3.0→4.0，4.0LE（2004.10.22）
- 「研究デザイン」の項を「論文の PECO を探る」から独立
- チェック項目毎に RCT の流れを図示
- 結果の項で，時間軸に垂直な指標と平行な指標について記述

4.0→4.1，4.1LE（2004.11.18）
- 追跡率の項の記載を変更
- 結果の項の記述を充実化

4.1→5.0，5.0LE（2005.11.12）
- レイアウトの変更
- 記述の探し方のガイドを新設

5.0→5.1，5.1LE（2005.12.3）
- ITT 解析と FAS の解説を変更
- PROBE 法の解説を追加

5.1→5.2，5.2LE（2006.3.10）
- 「介入以外の治療は同等か？」の質問項目を削除し，PECO の E/C と Baseline の比較に含めた
- 「追跡率／脱落率計算ワークシート」を設置

5.2→5.3，5.3LE（2007.8.26）
- 隠蔽化の説明を修正
- 脱落の定義を明示
- 「6. 症例数は十分か？」の項目を「結果に有意差がある」か否かでチェックするように変更

5.3→5.4，5.4LE（2008.5.2）
・PROBE 法の解説を修正
・p 値の解説を修正
・一部の項目で選択肢の順序を変更
5.4→5.5（2009.6.14）
・マスキングの項のキーワードとして，aware，open を
追加
・LE 版の廃止
・誤字修正
5.5→6.0（2010.10.31）
・PECO を PICO に変更
6.0→6.1（2012.10.8）
・隠蔽化の説明を変更（Cochrane Handbook に準拠）
・ITT 解析の解説を変更（Cochrane Handbook に準拠）
・追跡率の項で，追跡の状態を細分化
・CAT シートの citation の記載欄を変更
6.1→6.2（2013.2.7）
・隠蔽化の説明を変更

Critically Appraised Topic for Clinical Trial

Reviewer: 年 月 日

authors：

title：

citation：

PubMed PMID：

1. 論文の PICO は何か？

P：

I：

C：

O：

2. ランダム割付けされているか？

□ランダム □非ランダム 割付け方法：□中央割付け □封筒法 □その他（ ）

ランダム割付けが隠蔽化 concealment されているか？：□隠蔽化 □隠蔽化なし □不明

3. Baseline は同等か？

□差がない □差がある→どこに？（ ）

結果に影響を与える可能性のある因子は全て検討されているか？

○検討されている ○不足しているものがある→何？（ ）

4. 全ての患者の転帰が Outcome に反映されているか？

4-1. ITT 解析か？

□ITT □ITT でない→結果をくつがえしうるか？□くつがえしうる □くつがえさない

4-2. 結果に影響を及ぼすほどの脱落があるか？

□ない □ある 追跡率＝結果の n／割付時の n＝（ ） □不明

5. マスキング(盲検化)されているか？

マスキング（盲検化）されているのは誰か？

○患者，参加者 ○介入（治療）実施者 ○Outcome 評価者 ○データ解析者

□四重 □三重 □二重 □一重 □盲検なし □盲検化不可能 □不明

6. 症例数は十分か？

□結果に有意差がある →症例数は十分 →サンプルサイズは？○計算されている ○計算されていない

□結果に有意差がない →症例数は十分かどうか不明

→サンプルサイズは？ ○計算されている

研究に参加した人数は計算されたサンプルサイズを？

○超えている →症例数は十分 ○超えていない →症例数は不十分

○計算されていない

症例数（各群： 合計： ） イベント発生率： % 効果 % α： power：

□不明

7. 結果の評価

□時間軸に垂直な指標

追跡期間＝（ ）

介入群の発生率＝a／(a+b)＝（ %）＝EER

対照群の発生率＝c／(c+d)＝（ %）＝CER

RR＝EER／CER＝（ ）

RRR＝1－RR＝（ ）

ARR＝CER－EER＝（ ）

NNT＝1／ARR＝（ ）

その他の評価方法（outcome が連続変数の場合など）：

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

□時間軸に平行な指標（平均生存期間など，Kaplan-Meier 曲線）