GRADE*システムと SoF**

―エビデンスから推奨へ（Ⅱ）―

相　原　守　夫[1)]　Gordon Guyatt[2)]　Yngve Falck-Ytter[3)]

要　　約

背景：多くの診療ガイドラインが一般的になってきているが，エビデンスの質と推奨度の判定システムは非常に多種類があり，混乱のもとになっている。

目的と意義：GRADE（Grading of Recommendations Assessment, Development and Evaluation）システムは，コクラン共同計画，WHO，英国 NICE，UpToDate など，25 以上の機関で採用されており，われわれは本論文においてGRADE システムについて紹介する。

1）GRADE システムでは，推薦をグレーディングする際，患者の価値観や好みが，どう重要な役割を果たすかを説明している。

2）GRADE では，9 ポイントスケールを使用して，患者にとって重要なアウトカムを，重大，重大ではないが重要，重要でない，の三つのカテゴリーに分類する。

3）推奨度は，2 種類（強，弱）で，四つの因子（エビデンスの質，望ましい効果と望ましくない効果のバランス，価値観・好み，コスト）により決定される。

4）エビデンスの質（高，中，低，非常に低）は，研究デザインと，5 つのグレードダウン因子（研究の欠点，結果が不一致，エビデンスが間接的，不精確なデータ，報告バイアス）と，大きな推定効果，用量-反応勾配を含む 3 つのグレードアップ因子を考慮して判定する。

最近，GRADE ワーキンググループにより開発されたエビデンス要約（Summary of Findings：SoF）作成のための有用なツールが，コクラン共同計画の Review Manager（RevMan 5）と統合された。診療ガイドライン作成者や臨床医は，SoF 表を使った GRADE システムを理解することで，推奨決定や診療のメリットになるだろう。

はじめに

GRADE システムの世界的な普及が止まらない。継続的で，しかも疾患領域を横断しての広がりであり，BMJ 誌へのコクラン共同計画の発表[1)]から，多くの国でガイドライン作成あるいはその評価，利用のために GRADE システムが採用され始めている[2~4)]。

医療の最前線では，臨床医は毎日数多くの臨床的意思決定（clinical decision making）を行う。それらの多くは，患者の疑問に対しての決断で

Key words：GRADE system, Recommendation, Evidence, Patient-important outcome, Summary of findings

*GRADE：Grading of Recommendations Assessment, Development and Evaluation

**SoF：Summary of Findings

1) 相原内科医院　2) Department of Clinical Epidemiology and Biostatistics, HSC-2C12, McMaster University

3) Case Western Reserve University, School of Medicine, VA Medical Center Cleveland

あり，利益とリスク・害・負担のバランスを勘案し，患者の利益に最もかなうと判断する介入(治療や診断など)を推奨することにある。エビデンスを同定・収集するために，非系統的アプローチをとったのでは，偏った評価になりやすく，バイアスの可能性を限りなく減らすように努力したデザインによる研究手法で臨床疑問を扱ったシステマティックレビューやメタアナリシス，診療ガイドラインが参考となる。

診療ガイドラインは，「個々の臨床状況ごとに，医療者や患者が適切なヘルスケアを判断できるよう手助けするために系統的に作成された推奨」であり[5]，1970年代以来，さまざまな組織が，さまざまにエビデンスの質（レベル）と推奨の強さについてのグレーディングシステムを使ってきた。しかし，残念ながら，エビデンスレベルの質の等級付け方法や，推奨レベルのグレード付けの方法は，それを作成した組織によって種々に異なっている。同じエビデンスでも，作成組織が異なれば，異なる質のレベルに分類されうるため，当然ながら推奨グレードにも違いが出てくる。エビデンスの質レベルや，推奨グレードを表す際にも，記号（ABC …），番号（123，Ⅰ Ⅱ Ⅲ…），その組み合わせ（Ⅰa，Ⅰb，Ⅱa…），シンボルマークなどが用いられているが，この状況は効果的な情報伝達を妨げ，混乱をもたらしている[6〜17]。

優れた“エビデンスのレベルと推奨度”の評価基準とは，エビデンスの質と推奨度が別々に表示，ガイドラインの消費者である臨床家にとって記載が簡便明瞭，ガイドライン作成の方法論が明示，ガイドラインの作成が簡便，適切な数のカテゴリー，グレーディング・システムに一貫性があるなど，推奨のもとになる基本的な価値判断を明示し，アウトカムや患者因子を考慮すべきものである[9,18,19]。

患者中心の医療において，医師の専門性・経験・利用可能技術，患者側因子，エビデンスの質，の3要素を統合して判定することが重要であり，現状における推奨度やエビデンスの質の判定基準の国際的な不均一性を解消すべく登場したのが，Grading of Recommendations Assessment, Development and Evaluation（GRADE）ワーキンググループによるGRADEシステムである[1,2]。

コクラン共同計画をはじめとした海外でのGRADEシステムの普及，診療ガイドライン作成・評価への本システムの採用と比較すると，国内では認識度そのものが非常に低く，わずかに本システムが参考・参照されている国内診療ガイドラインがあるものの，本システムの適切な理解による利用とはいえない。われわれは2007年1月より，国内でのGRADEシステムの迅　速，かつ適正な理解・普及を進めており，最近GRADEシステムについての紹介報告を行った[20]。今回は，さらにコクランREGISTERED METHODS GROUPSによる本システムでのエビデンス要約表（Summary of Findings［SoF］table），コクランReview manager（RevMan 5）との統合性報告[4,21,22]，あるいは*Cochrane Database of Systematic Reviews*でのGRADEシステムに準じたSoFを使用した発表[23]，ワーキンググループによるFAQ，key論文[1,4,24]などを参考にして紹介する。

本システムで多用されるいくつかの用語の意味は，末尾の用語集（**表7**）を参考にしていただきたい。

Ⅰ　GRADEとは

GRADEワーキンググループは，医療の推奨度の作成・評価において，既存のグレーディングシステムの短所を改善したいと願う人々により，Oxmanのリードのもとに非公式の共同体として2000年に発足した。GRADEシステムはこのワーキンググループにより2004年に発表され[1,2]，アウトカムの重要性を主点として，推奨度を強弱の2種類とし，全体的なエビデンスの質を4段階カテゴリー（高，中，低，非常に低）としている。本システムは，単に診療ガイドラインを利用，あるいは作成する側だけのためではなく，診療ガイドラインの根幹である“推奨”について，「つくる」，「つたえる」，「つかう」の

表 1　推奨度の 2 分類（文献 24，25）

1．強い推奨	ある介入による望ましい効果（利益）が，明らかに（clearly），望ましくない効果（downsides：害・負担・コスト）をしのぐ，あるいは明らかにその逆である場合。
（表現例）	“we recommend for〜”or“recommend against〜” “clinicians should〜”or“clinicians should not〜”
2．弱い推奨	(1) 利益/downsides（害・負担・コスト）のバランスが不確か（less certain）。 (2) エビデンスの質が低い。 (3) 利益と downsides が比較的接近していることを示唆する“高い質”のエビデンスがある。
（表現例）	“we suggest〜”or“suggest not〜” “clinicians might〜”or“clinicians might not〜”

表 2　ガイドライン利用者（患者・臨床医・行政）にとっての「強い推奨」「弱い推奨」の意味（文献 25）

〔強い推奨〕
- 患者に対して：この状況ではほぼ全員がその推奨に沿った診療を希望し，ほんの一部の人たちが受け入れないだけだろう。
- 臨床医に対して：ほとんどの患者がその介入を受け入れられるようにすべきである。
- 行政に対して：ほとんどの状況で施策として採用することができる。

例：深部静脈血栓症患者に対する肺塞栓予防のため抗凝固療法を早期に開始。
　　市中肺炎に対する抗生物質使用。手術後感染予防のための手術直前 1 回抗生物質使用。

〔弱い推奨〕
- 患者に対して：この状況では，半数以上の患者は示唆された診療方法を希望するはずだが，その診療方法を希望しない患者も多い。
- 臨床医に対して：患者の価値観や選好に沿う意思決定を行うための意思決定支援が有用であろう。エビデンスやエビデンスの要約を自分自身で再検討すること。
- 行政に対して：行政的には，多くの利害関係者の参加のもとで，かなりの議論が必要であろう。

例：運動能力の低下した（上葉優位の）重症肺気腫患者に対する肺容量減少手術。
　　特発性静脈血栓塞栓症（VTE）患者に対する継続的（漫然とした）抗凝固療法。

三者にとって，単純かつ明瞭なシステムである。

また，エビデンスの要約（SoF）作成支援のソフト（GRADE *profiler-β*）の開発により，コクラン共同計画での RevMan 5 によるレビュー要約との統合によるユーザーの理解度の向上につながっている[4,21,22]。

Ⅱ　GRADE システムでの推奨度

本システムでの推奨の程度（推奨度）は，強（strong），弱（weak）の 2 種類である（**表 1**）[24,25]。また，推奨の方向性も，介入を推奨する（recommend for），しない（recommend against）の 2 種類があり，結果としての推奨表現としては 4 種類となるが，GRADE システムでは中間層の推奨度分類は作成しないようにしている。推奨度の分類が 2 種類で少ないのは，臨床医，患者，政策担当者の理解を容易にするためであり，GRADE ワーキンググループでは，あえて各立場の違いによる推奨度の意味を具体的に説明している（**表 2**）[1,24,25]。

推奨度とは，“介入による望ましい効果（利益）が，望ましくない効果，downsides（害・負担・コスト）を上回るか（positive），あるいはその逆か（negative）について，どの程度確信できるかを示すもの”と定義され，推奨の強さとエビデンスの質とは分離している。望ましい効果

表 3　推奨度に影響する 4 つの主要因子，強い推奨，弱い推奨の例（文献 20，26）

	強い推奨例	弱い推奨例
エビデンスの質	多くの質の高いランダム化比較試験が気管支喘息での吸入ステロイドの利益を証明している。	気胸に対する胸膜癒着の効用を検討しているものは症例集積研究が一つあるだけである。
望ましい効果と downsides のバランス	心筋梗塞でのアスピリンは死亡率を減少させ，毒性や不便さ，コストも少ない。	低リスクの心房細動でのワルファリンはわずかに脳卒中減少をもたらすが，出血リスクの上昇や，かなりの不便さがある。
価値感や好みにおける多様性	悪性リンパ腫の若年患者では，その多くが，治療に伴う毒性より，化学療法による寿命延長効果の方に高い価値をおくだろう。	高齢の悪性リンパ腫患者では，治療に伴う毒性よりも，化学療法による寿命延長効果のほうに高い価値を置くことはできないかもしれない。
介入が医療資源の賢明な使用か否か	一過性脳虚血発作の患者での脳卒中予防としてのアスピリンは，低コストである。	一過性脳虚血発作の患者での脳卒中予防としてのクロピドグレルやジピリダモール／アスピリンはコストが高い。

とは，死亡率の低下，特定の疾患による死亡率低下，罹患率低下，QOL 改善，治療の負担軽減（治療を受ける必要性，血液検査やモニターのための通院の不便性など），医療費の減少であり，望ましくない結果とは死亡率上昇，罹患率上昇，あるいは QOL への悪影響，害反応症状や検査異常などである。

推奨度の強さは，その定義からも明らかなように質の高いエビデンスは強い推奨につながることが一般であるが，推奨度は臨床に直結したものであり，エビデンスの質を含め考慮すべき四つの主要因子がある（**表 3**）[20,26]。

1　推奨度決定の 4 主要因子

推奨の程度（推奨度）を決定するための 4 つの主な因子がある。

（1）エビデンスの質

推奨度の決定因子の一つがエビデンスの質であり，エビデンスの質が高いほど，強い推奨につながる。観察研究と実験的な研究デザイン（ランダム化比較試験：RCT）とでは，研究に用いる群の予後が異なっているが，観察研究ではブラインド化のようなバイアスを避けることが難しく，一般的に観察研究によるエビデンスは，RCT によるエビデンスよりも，エビデンスの質としてはかなり低くなる。

（2）望ましい効果と望ましくない影響とのバランス

利益と downsides（害・負担・コスト）とのバランスで，利益が downsides を明らかに上回る場合や，逆に downsides が利益を明らかに上回る場合には，その推奨の確信は強い（strong）ものになる。利益と downsides が拮抗している場合（trade-offs の関係）や，利益と downsides，あるいは両者のいずれかの程度に不確かさがある場合は，推奨度は弱い（weak）ものになる。たとえば，低リスク心房細動患者でのワルファリン治療などである。

（3）価値観と好み

患者の価値観や好みが多彩なほど，あるいは不確実なほど，推奨の程度は弱くなる。特発性深部静脈血栓症（DVT）の 40 歳男性患者で，1 年間用量調整ワルファリン治療を受けている例を考えてみよう。もしその患者が標準強度のワルファリン治療を継続するなら，DVT 再発は約 10%/年の減少となることが推定される。毎日のワルファリン錠服用やビタミン K 摂取量を一定にすること，併用薬剤の注意，抗凝固治療モニターの血液検査，など治療による負担と，

表 4　エビデンスの質と推奨度の表現の例（文献 15）

エビデンスの質	シンボル	文字
高	⊕ ⊕ ⊕ ⊕	A
中	⊕ ⊕ ⊕ ◯	B
低	⊕ ⊕ ◯ ◯	C
非常に低	⊕ ◯ ◯ ◯	D
推奨度	**シンボル**	**番号**
強い推奨（介入の実行あるいは反対）	↑↑ OR ↓↓	1
弱い推奨（介入を実行あるいは反対）	↑？ OR ↓？	2

GRADE では，特に上記のいずれを使用することを規制しているものではないが，文字や番号にこだわるならば，推奨度には番号を，エビデンスの質の判定表示には文字を使用するよう薦めている。

出血リスクを比較して，DVT 再発を強く嫌う患者は，それらの downsides はワルファリン服用に値するものと思うかもしれないし，患者によっては，その利益はリスクに見合ったものでないと考える人もいるだろう[13]。また，若年者と高齢者では，延命に対する価値観は同じではなく，治療による利益と downsides に対する評価は異なりうる[25]。

（4）コスト

より高いコスト，あるいは医療資源のより大きな消費の介入は，推奨度は弱いものになる。一過性脳虚血発作患者では，アスピリン投与よりクロピドグレル投与などがコストはかなり高く，より高いコストは介入の強い推奨を低下させることになる[25]。

医療の利害得失を比較する際，アウトカムの一つとしてコストを考慮すべきであるが，費用は他のアウトカムより，時間，地理的区域，および含意の上ではるかに多彩である。特許期限が切れると薬剤コストは急に下がる傾向があり，同じ薬剤に対する費用も国が異なり，規制が異なれば著しく異なってくる。さらに，医療資源もかなりばらつきが大きく，たとえば，同じ高価な薬剤の年間処方費用は，米国の独身看護師 1 人の給料，ポーランドの看護師 6 人分，および中国の看護師 30 人分の給料を支払うことに相当している。コストの影響を強く受ける推奨は，医療資源の拡大に応じて，また時間の経過により変化しやすいものである[18,26]。

2　推奨度の用語・表現

推奨度を表現する用語としては，“強・弱”にこだわる必要はなく，たとえば“推奨する（recommend）”や“勧める（suggest）”，あるいは，“～すべき（should）”や“～してもよい（may, might）”など，実際にはさまざまな表現が用いられている。ただ，従来記載されていた，do it, probably do it, probably don’t it, don’t it は，誤解が生じやすく好ましくないという会議内意見から現在の表現になっているが，これらの用語を禁止しているものではない。また，後述のように，推奨度の表現方法として，シンボルや番号の使用も許容している（**表 4**）[15]。

3　推奨度と重要性

推奨の強さとその重要性は必ずしも同じではなく，強い推奨でも，患者の視点から（より一般的には医療制度の観点から）考えると，重要ではない場合がある。推奨を選択することによって生じる影響がそれほど重要ではない場合，強い推奨であっても気にかけない患者がいる。これは，患者が多種類の新薬や，生活スタイル改善のための多くの提案を与えられた際に特によく起こる。

政府や公衆衛生局は，推奨の強さだけでなく多くの問題を考慮しなければならないため，各

患者にとっては重要な強い推奨であっても優先度が低いと考えることがある。一般的に臨床医からの推奨とはほとんど関連性のないこれらの問題として，健康問題の蔓延（より蔓延している状態ほど優先度は高い），公平さの考慮（不利な条件に置かれた人々を対象に，健康上の公平さに取り組む治療の優先度は高い），社会に与える総医療費（総医療費が非常に高い治療の優先度は低い），医療の質改善の可能性（利用されていない治療の優先度は高い，など）が挙げられる。したがって，ガイドライン委員会は，資金提供者または医療制度管理者に対して，健康問題の蔓延，公平さ，医療の負担，医療の質改善の問題が推奨に及ぼす影響を透明化する必要がある[26]。

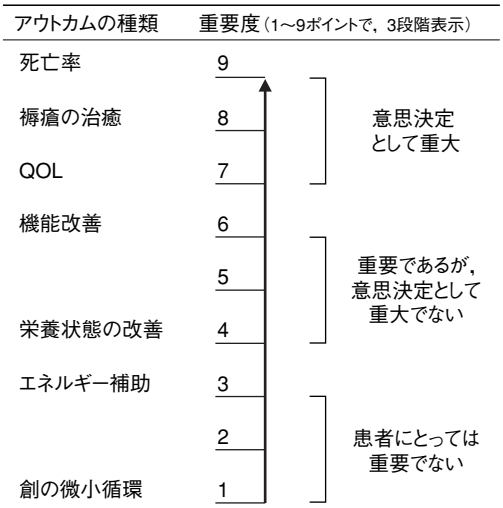

図 1　患者の重要度に従ったアウトカムの階層表示例――褥瘡を有する老人病患者での経腸栄養の効果（文献 27）

Ⅲ　GRADE システムにおけるアウトカムの 3 段階評価

1　GRADE によるアウトカム

GRADE システムでは，各アウトカムを 1〜9 ポイントで相対的に判定し，さらに 3 分画に分ける。GRADE システムでは，アウトカムを抽出する際，患者にとっての各アウトカムの重要性を相対的に判断し，重大（critical），重要だが重大ではない（important but not critical），重要ではない（not important）の 3 段階とする。最終決定に際して，各研究での著者が主要アウトカムとしているものや，他のアウトカム，さらには自分がなんらかの決定を行う上で重要かもしれないと思う（利益と害，関連したものならコストまで含め）他のアウトカムにも留意するが，全アウトカムのうち重大・重要なものをエビデンス・プロフィール評価の対象とする。アウトカムが重大か，重大ではないが重要か，重要でないかの決定は，一つの価値観であり，その際には，提案する推奨が影響を及ぼす人の価値観をできるだけ考慮すべきである。

図 1 はアウトカムの階層構造を示すが，褥瘡を有する老人患者での経腸栄養効果についてのエビデンスを調査するためのアウトカムの相対的重要性評価のアプローチ例である[27]。レビューやガイドラインの作成者はスケールの上段（7〜9）に相応するものを重大なアウトカム，中段スケール（4〜6）を重要だが重大ではない，下段スケール（1〜3）は臨床研究でときどき測定される代理アウトカムなどで，患者には直接的な重要性はないものとする。ガイドラインパネルは，この例が示すような明白なアプローチの種類を求めて努力すべきであり，コクラン・レビューと同様に重要なアウトカムを重視してワークシートにまとめるべきと報告されている[28]。

どのような診療ガイドラインや推奨の作成においても，介入（治療や診断など）に関しての臨床疑問（clinical question：CQ）を設定し定式化するが，それは 4 つの因子から成り立っている。すなわち，Patient（population），Intervention，Comparison，Outcome である。最近報告された膵臓癌例での外科治療論文（幽門輪保存法と標準 Whipple 膵頭十二指腸切除術を比較した 6 件の RCT を統合したメタアナリシス研究，n＝574）[29]を例にすると，膵臓癌で手術した患者で（P），幽門輪残存の改変手術は（I），標準の広範囲切除術と比較し（C），短期・長期の死亡率，

表 5　GRADE エビデンス・プロフィールの例（文献 29）

CQ：膵臓癌あるいは乳頭部癌患者で，幽門輪残存手術の効果（PPPD）を，標準 Whipple 膵頭十二指腸切除術（SWPD）と比較する。

P（patient or population）：手術可能な膵臓癌あるいは乳頭部癌，I：pylorus-preserving pancreaticoduodenectomy，C：standard Whipple pancreaticoduodenectomy

settings：入院患者

研究デザイン：ランダム化比較試験

Outcome measure	質の評価（Quality assessment）						研究結果のまとめ（Summary of findings）			
アウトカム	研究数 参加数	研究の欠点	一致性	直接性	精確性	報告バイアス	RR[1] （95% CI）	最良平均値 （SWPD 群）	ARR[2] （95% CI）	エビデンスの質
5 年死亡率	3 229	深刻な欠点（−1）	重要な問題なし	問題なし	精確	ありそうでない	RR 0.98 （0.87〜1.11）	82.50%	−0.02 （−0.12〜0.08）	＋＋＋ 中
周術期死亡率	6 490	深刻な欠点（−1）[3]	重要な問題なし	問題なし	不精確（−1）[4]	ありそうでない	RR 0.40 （0.14〜1.13）	4.90%	−0.02 （−0.05〜0.01）	＋＋ 低
輸血（単位）	5 320	深刻な欠点（−1）[3]	重要な問題なし	問題なし	不精確（−1）[4]	ありそうでない	—	2.45	WMD：−0.66 （−1.06〜−0.25）	＋＋＋ 中
胆汁漏出	3 268	深刻な欠点（−1）[3]	重要な問題なし	問題なし	不精確（−1）[4]	ありそうでない	RR 4.77 （0.23〜97.96）	0.00%	0.02 （−0.02〜0.05）	＋＋ 低
入院期間（日）	5 446	深刻な欠点（−1）[3]	重要な問題なし	問題なし	不精確（−1）[4]	ありそうでない	—	19.17	WMD：−1.45 （−3.28〜0.38）	＋＋ 低
胃内容排泄遅延	5 442	深刻な欠点（−1）[3]	説明不能な不一致（異質性）（−1）[5]	問題なし	不精確（−1）[4]	ありそうでない	RR 1.52 （0.74〜3.14）	25.50%	0.11 （−0.08〜0.29）	＋ 非常に低

[1]：RR＝相対危険（ランダム効果モデル）　[2]：ARR＝絶対リスク減少率；WMD＝重み付け平均差　[3]：全研究にわたり割付の隠蔽が不確実，1 件の研究で患者盲検化，アウトカム評価のブラインド化なし，3 研究で脱落率（＞20%），1 研究で intention-to-treat 解析なし　[4]：信頼区間の間に 1 があるため，どちらの手術方法の利益（あるいは害）が大きいか不明　[5]：異質性：I^2＝72.6%，p＝0.006

輸血，胆汁漏出，入院期間，胃内容排泄遅延についての効果・影響（O）である（**表 5**）。

2　一般的なアウトカム

通常，アウトカム（エンドポイント）には，2 値アウトカム（脳卒中，心筋梗塞，死亡などのイベント発生の割合など）と，連続値アウトカム（潰瘍の症状の減少や日数，輸血量，肺機能の変化など）の 2 種類がある。前者では，オッズ比（OR）や相対危険（RR）がよく使われるが，後者では，測定尺度の単位が同じ場合は介入群と対照群の平均値の差（加重平均の差 weighted mean difference：WMD）を使って統合し，測定尺度の単位が異なる場合には，標準化平均差（standardized mean difference：SMD），すなわち，平均の差を標準偏差で除したもので統合する。アウトカムが 2 値でない場合でも，なんらかの閾値を設定することで 2 値変数に変換は可能であるが，当然ながら情報量の低下を伴うものである。

また，RR と OR との間には，イベント発症率が低い場合（多くのランダム化比較試験でそうであるように），両者は非常に近い値となる。しかし，イベント発生率が高く，効果サイズが大きい場合でも，OR か RR に換算する方法がある。さらに，RR はコントロール群でのイベント発生率が変化しても全く影響がないが，絶対危険減少（ARR）はベースラインリスクが変化すると著変する。同じように，相対危険減少率（RRR）や治療効果発現必要例数（NNT）も同様に変化し，これは副作用発現必要治療症例数（NNH）でも同様である。このようにイベント発生率が低い場合は，RR は OR と極めて近いと仮定することができるが，リスクが高いほど仮定は疑わしいものになる。それでは，どの関連性の指標が最もよいのだろうか？　最もよい選択は，2×2 表もしくは生存分析の形でも，相対的数値も絶対的数値もともに表示することである。結果を吟味すると，NNT の算出が最も役立つことになるが，NNT だけでの表示では効果が低く評価されやすいことは以前から報告されており，これは製薬会社が治療効果の説明に際して相対的指標を多用する一因であろう。同じ現象は，臨床医と患者の間でも起こりうることで，NNT を使って表現した場合のほうが，効果が低く評価されたと報告されている[9,30]。

Ⅳ　GRADE システムにおけるエビデンスの質

GRADE システムでのエビデンスの質は 4 種類（高，中，低，非常に低）に判定され，本システムでは，集めた複数の論文より，種々のレベルのアウトカム（重大，重大ではないが重要）ごとに，エビデンスの質を評価する。エビデンスの質とは，“推定効果（estimate of effect）に対する確信（confidence）のもてる程度”と定義される。重要なアウトカムに関する複数の研究を通してのエビデンスの質は，コクラン・レビューのようなシステマティックレビューなどを用いて判断することが可能であるが，後述のバイアスなどにも留意すべきであり，また概してレビューから得られる結果以上の情報が必要である[1]。

GRADE システムでは，明瞭かつ単純に評価するために，最初に，研究デザインによってエビデンスの質を分類し，RCT は“高（high）”，観察研究は“低（low）”，その他のエビデンスを“非常に低（very low）”とする（**表 6**）[1,24]。

観察研究によるエビデンスの質は，一般的に RCT によるエビデンスの質よりも低くなり，その顕著な例としては，ホルモン補充療法による冠動脈心疾患のリスク評価がある。観察研究ではリスク低減効果が認められたが，その後の RCT ではリスク低下は認められず，むしろリスクが増大していた[31]。また，研究デザインでは RCT で高いレベルに分類されても，最低ランクに分類されることがあり，症例対照研究のような低いレベルでも最終的には高いレベルに分類されることがありうる。

このように最初の時点では，研究デザイン評価として，“moderate（中）”というカテゴリーはなく，また“非常に低”の研究に対しては，グレードダウン・アップの評価はしない。最終

表 6　エビデンスのグレーディングの方法（文献 25，FAQ）

エビデンスの種類	①ランダム化比較試験＝高 ②観察研究＝低 ③その他のエビデンス＝非常に低
グレードダウンの 5 項目	①RCT の方法にかかわる欠点（limitations）：RCT のデザインおよび実施結果に“深刻”な欠点がある。（−1）*，“非常に深刻”な欠点がある。（−2）** ②結果が不一致（inconsistency），または異質性（heterogeneity）がある（サブ解析の問題も含め）。（−1）* ③患者への適用上，エビデンスが間接的（indirect）：対象集団や介入方法，コントロール，アウトカムの直接性が，乏しい（−1）*，あるいは，極めて乏しい。（−2）** ④データが不精確（imprecise），あるいは，まばら（sparse）である。（−1）* ⑤報告バイアスが存在する可能性が高い。（−1）*
グレードアップの 3 項目	①・強い関連性（association），大きな効果・影響（magnitude of effect）が示される（交絡因子の可能性のない，複数の観察研究が一致している）。すなわち，相対リスク（RR）＞2 あるいは RR＜0.5。（＋1）* ・妥当性が脅かされるおそれがなく，極めて強い関連性（もしくは効果・影響の程度が極めて大きい）がある。すなわち，RR＞5 あるいは RR＜0.2。（＋2）** ②用量-反応関係（勾配）がある。（＋1）* ③交絡因子の可能性があるが，そのすべてが結果に対して，効果・影響を減少させる方向に働いたと考えられる場合。（＋1）*
エビデンスのグレード	①高＝さらなる研究を実施しても，推定効果への確信は変わることはほとんどない（very unlikely）。 ②中＝さらなる研究が，推定効果への確信に重要なインパクトをもつ可能性があり（likely），結果としてその推定が変わるかもしれない（may）。 ③低＝さらなる研究が，推定効果への確信に重要なインパクトをもつ可能性があり（very likely），結果としてその推定が変わる可能性がある（likely）。 ④非常に低＝どの推定効果も非常に不確かなものである（very uncertain）。

カッコ内の数値は，エビデンスの質の変化を示している。
（−1）*，（＋1）*＝1 段階グレードがアップ・ダウン（例：高から中へ，あるいは低から中へ）
（−2）**，（＋2）**＝2 段階グレードがアップ・ダウン（例：高から低へ，あるいは低から高へ）

的な 4 分類（高，中，低，非常に低）の各カテゴリーの意味は，高＝効果がどの程度確実なのかという推定が，さらに研究を重ねても変わる可能性が少ないもの，中＝さらに研究を重ねた場合，効果がどの程度確実なのかという推定に重大な影響を与え，推定が変わる可能性のあるもの，低＝さらに研究を重ねた場合，効果がどの程度確実なのかという推定に重大な影響を与え，推定が変わる可能性が高いもの，非常に低＝推定される効果がまったく不確実なもの，である。

また，いわゆる“エキスパートの意見”をエビデンスのカテゴリーに分類する際には，やや混乱がある。もし，それがエキスパートにより，すべてのエビデンスを検討して質が高いか低いかを評価し“判定”されたものなら，エキスパートの“判定”は適切（高，あるいは低）なものとしてよい。しかし，“エキスパートの意見”が，エキスパートの報告や彼らの経験という意味ならば，それらは症例報告や，対照を置かない臨床観察と同様（あるいはそれ以下の），“非常に低”のレベルとすべきである。

さらに，研究デザインとは別に，エビデンスの質は，多くの因子で変化することがある。それらの主要因子としては，グレードダウンの 5 項目と，グレードアップの 3 項目である（**表 6**）。

1　グレードダウンの 5 因子

エビデンスの質の低下の代表的な因子として，研究の欠点（study limitations），結果の不一致（inconsistency），エビデンスの間接性

(indirectness)，データの不精確（imprecision），報告バイアス（出版バイアスを含む）がある。

（1）研究の欠点（study limitations）

研究計画そのものと，実施した研究全般についての欠点(問題点)，risk of bias のことである。RCT については，割付の隠蔽がないこと，バイアスの影響を受けやすい主観的なアウトカムを用いた研究，ブラインド化されていないもの，フォローアップでの大きな脱落，ITT（intention-to-treat）解析を遵守していないもの，利益が見込まれたとして試験を早期中止した場合，などである。早期中止については，今後，群逐次手法を用いた研究が増える可能性もあり，注意が必要である。これらの研究上の欠点（limitations）の程度は，深刻か，非常に深刻か，の2種類で，おのおの1段階，2段階グレードダウンとする。

また，膵臓癌手術例[29]では，割付遮蔽の欠如，患者やアウトカム評価者のブラインド化の欠如，フォローアップの脱落などの欠点があり，5年死亡率，輸血量という，おのおのの重要なアウトカムについてのエビデンスの質は，“高(high)”ではなく“中（moderate)”であった。さらには，周術期死亡率，胆汁漏出，入院期間というアウトカムについての評価は“低（low)”であり，後述のように，胃内容排出遅延については“非常に低（very low)”であった（**表 5**)。

（2）結果が不一致（inconsistency of results）

いくつかの試験で得られた結果が相互に一致せず（異質性 heterogeneity，あるいは多様性 variability)，その理由が説明できない場合には，結果の不一致があるとする。多様性 variability は，集団の違い（例：薬剤によっては，より病気の重い対象集団に投与された場合では相対的効果が大きくなる)，介入の違い（例：高用量の薬剤はより効果が大きくなりやすい)，アウトカムの相違（例：治療効果が経過とともに減弱する)，などでも起こりうる。もし，各研究結果が一致せず（異質性あり)，研究者が納得できる説明ができないなら，エビデンスの質は低下することになる。この inconsistency の存在は，1段階グレードダウンとする。

膵臓癌手術例[29]では，胃内容排泄遅延（delayed gastric emptying）をアウトカムとした幽門輪保存術の標準 Whipple 膵頭十二指腸切除術との比較では，方法論上の深刻な欠点によって低下して“中”のレベルとなったエビデンスの質を，異質性（I^2値 72.6%，$p=0.006$）[32]の存在のためにさらに“非常に低”まで低下させることになった（**表 5**)。

（3）エビデンスの間接性（indirect）

推奨度の判定・作成者はエビデンスの直接性(directness)に関して2種類の問題に直面する。一つは，2種の薬剤（A と B）のうち一つを考慮する場合で，A と B を比較したランダム試験がないが，A とプラセボ，B とプラセボのランダム化比較試験がある場合である。このような場合，A と B の効果の程度を間接的に比較することになり，直接比較よりもエビデンスの質は低下する。第2のタイプは，対象患者，介入，比較対照，アウトカムについての相違であり，臨床試験の測定対象が，どの程度，関心の対象に似ているかである。もし目の前の患者が，研究内容と比較して，より高齢，重症，合併症が多く，異なる場合には直接性は不確実なものになり，間接的（indirect）と評価される。また，治療介入例では，薬剤の種類が同クラスでも異なる場合，用量，使用回数の違い，などで直接性が不確かとなる。さらに，代理アウトカムを用いる研究は，重要なアウトカムを用いる研究と比較すると，エビデンスの直接性が十分ではない。この直接性（directness）が，ある程度不確実ならば1段階，かなり不確実ならば2段階グレードダウンとする。

（4）不精確（imprecision）

標本サイズが小さい場合やイベント数が少ない研究は，ランダム誤差が生じやすく，信頼区間が広くなり，利益あるいは，downsides（害，負担，コスト）についても，その結果は不確実とすべきである。また，研究が1件しかない場合にも，不精確と判定すべきであり，1段階レベルダウンとなる。

膵臓癌手術例[29]での幽門輪保存の代替手術の多くのアウトカムでは，著明効果と無効の両者の存在や，あるアウトカムでは両方向での重要な相違が存在している。また，前例の胃内容排泄遅延のアウトカムにおいては，RR＝1.52（95％信頼区間 0.74〜3.14）であり，最終的には，エビデンスの質は“非常に低”までに低下することになった（**表 5**）。

（5）報告バイアス

バイアスとは，真実からの一定方向への偏りをもたらすものである。たとえば，検討対象としている薬剤に有利な結果が得られれば出版されやすく，有益性が認められないとか，害が認められた場合には出版されることが少なく，これは出版バイアスと呼ばれ，報告バイアスの最たるものである。また，たとえ研究結果が出版されたとしても，アウトカムのうち検討対象の薬剤に都合のよいものだけが報告され，都合の悪いアウトカムの結果が報告されないことがしばしばある。極端な場合は，総死亡が増加傾向を示しているのに，ある疾患の罹患率が低下したことのみで有効としているような例がある。このように報告バイアスの可能性が高い場合も，1 段階グレードダウンにつながる。

コクラン・レビューにおいてさえも，同定された RCT の約半数が患者にとって重要と思われるエンドポイント（アウトカム）の効果推定に貢献できていない。またメタアナリシスで同定された RCT のうち実際に効果推定に貢献できた試験の占める割合と推定効果には逆相関がみられたことから，この割合の低いメタアナリシスは真実の効果を過大評価している可能性がある。これを outcome reporting bias（アウトカム報告バイアス）というが[33]，今まで注目されてきた study reporting bias に加えて，このアウトカム報告バイアスにも留意すべきである。

以上のような研究の欠点（limitations）が多ければ多いほど，エビデンスの質は低いものになる。すなわち，膵臓癌例[29]では，5 つのランダム化比較試験が利用できたとしても，各アウトカムによっては，最終的なエビデンスの質は“中”，ないし，“非常に低”にまで低下していた（**表 5**）。

2　グレードアップの 3 因子

良質であっても観察研究は一般にエビデンスの質は“低”であるが，まれながらエビデンスの質がグレードアップし，“中”，あるいは“高”のエビデンスと評価されることがありうる。主に三つの因子があり，①介入の効果の程度（magnitude of effect）が大きい，②交絡因子のために効果・影響が過小評価になっていると考えられる，③用量-反応関係（勾配）の存在，である（**表 6**）。

（1）効果の程度が大きい

治療効果や危険度などプールされたエフェクト・サイズ（effect size）が大きい場合である。RR＞2，あるいは RR＜0.5 の場合には効果・影響の幅が“大きく”，1 段階グレードアップとなる。また，複数の観察研究で，RR＞5 あるいは RR＜0.2 の場合には，効果・影響が“非常に大きい”と考えるのが妥当であろうとして，2 段階アップとされる。

たとえば観察研究のメタアナリシスで，自転車用ヘルメットはかなりクラッシュにかかわる自転車乗りの頭部外傷の危険を減少させると報告されており（OR 0.31，95％CI 0.26〜0.37）[34]，この大きな効果はエビデンスの質を“中”にグレードアップさせる。また，心弁膜置換術実施患者でのワルファリン予防治療効果を評価した観察研究でのメタアナリシスで，血栓塞栓症の RR は 0.17（95％CI 0.13〜0.24）[35]で，この“非常に大きな”効果は，エビデンスの質を“高”のレベルと示唆するものである。

（2）交絡因子のため，効果・影響が過小評価になっている

交絡因子のため，結果データで示された効果・影響が真の効果・影響よりも弱められていると考えられる場合であり，一段階レベルアップとする。もし，より重病の患者だけが実験的な介入治療を受けていて，それでも実験群でよりよい結果であれば，それは実際の介入治療効果は得られた結果以上の効果を示唆すると考え

られる。また，介入試験において，より重症の患者だけが治療を受けていて，それでもなお有効な結果であれば，それは実質的な治療効果は研究データよりも大きなものであろう。

(3) 用量-反応関係（勾配）の存在

たとえば，ワルファリンによる抗凝固療法を受けている患者で，INR のレベル上昇と出血リスクの増加の間には用量-反応勾配があることが観察研究で判明しており，治療域以上の抗凝固療法は出血リスクを増加させるという確信になっている。

3　アウトカムと全体的なエビデンスの質，SoF

以上のような因子を考慮して，最終的に各アウトカムについての全体的なエビデンスの質を決定し，4 つのカテゴリーのいずれかに判定する。

各アウトカムについての全体的エビデンスのレベルの質を等級するために，ガイドライン作成者は，多数の研究からのすべてのエビデンスについて，研究欠点などをまとめ，考慮する必要がある。既存の系統的レビューが非常に参考になるが，残念ながら系統的レビューでは通常はすべての重要なアウトカムについて記載していることはなく，多くは利益に関してのものである。

なんらかの推奨を判定する場合で，全体的なエビデンスの質を判定する上で重要なことは，どのようなアウトカムが患者にとって重大（critical）なのかということと，複数の重大なアウトカムにわたってエビデンスの質が異なる場合である。その際の原則は，すべてのアウトカムが同じ方向ならば問題はなく，この場合は最も“質の高いエビデンス”を採用するが，各重大なアウトカムが利益や害に関して別々の方向を示している場合，“全体的なエビデンスの質”は“重大（critical）なアウトカム”の“最も低いエビデンスの質”を採用する，である。

膵臓癌手術例[27]にたとえると，推奨を作成する者が，仮に胃内容排泄遅延が重大（critical）なアウトカムと判断するなら，全体的エビデンスの質は“非常に低”となるだろう。胃内容排泄遅延は重要だが重大とはいえない（not critical）ものならば，5 年死亡率に関するエビデンスの“中”の質があるにもかかわらず，(明らかに重大（critical）な周術期死亡率についての結果を基にして)，全体的エビデンスの質は“低”であろう。

忙しい臨床医，忙しい患者や行政担当者には簡潔，明瞭，容易に理解できる SoF table が有用であり[22]，GRADE システムではこのような支援のためのツール（GRADE *profiler*）を開発した[4,21]。これは，visual basic を基本とした spread sheet で，一連の CQ から，アウトカムの相対的なランキング，エビデンスの階層的判定，推定効果の算出などを支援する簡便なツールである。最近，RevMan 5 との統合・連携が可能となり，コクラン共同計画でも，GRADE Evidence table として最近発表され[23]，Cochrane Handbook（ver. 5）でも，2007 年 10〜11 月，RevMan 5 公開とともに，新しい“risk of bias tool”（allocation concealment が追加）が発表されている[22]。

V　GRADE システムでの推奨とエビデンスの質表示

医療の最前線においては，診断や治療の最適な医療行為としては，推奨度に最も興味があることから，本 system では，第一に“推奨度”を，次に“エビデンスの質”を表示する（例：強い推奨/エビデンスの質は“高（high）”）が，それらの利用例はすでに治療や診断についての論文[35]でもみることができる。また，GRADE システムでは，ガイドラインなどでも，推奨とエビデンスの質を，一見して内容が理解できるように，オリジナルのシンボル・記号や推奨度ナンバーリング（1，2），エビデンスの質（A，B，C，D）の利用も許容している。注意すべき点としては，文字やナンバーを使用する際には，推奨度にナンバーを，エビデンスの質に文字を使用した組み合わせにするよう勧めており[15]，それらの報告例もある（**表 4**）[19,26,36]。

Ⅵ　GRADE システムの現状と今後

本 GRADE システムは，従来の診療ガイドラインの推奨度作成あるいは臨床医の clinical decision making での推奨決定システムとは全く異なっている[1]（FAQ 参照）。アウトカムの重要性を評価し，各アウトカムについてのエビデンスの質を判定し，患者因子（価値観，負担，コストなど）を考慮して，その作成プロセスもわかりやすくした“推奨度”の判定作成方法である。

1　海外における GRADE システムの急速普及

GRADE ワーキンググループは，常にシステムのさらなる単純化，明瞭化をめざし，年間数回の会合を継続している。その単純性，明瞭性は海外での採用状況をみることでも容易に理解できる。Cochrane Collaboration[3,21,22]，American College of Chest Physicians（ACCP）[19]，American Thoracic Society（ATS）[24]，UpToDate[37]，WHO[38]，Ontario Ministry of Health and Long-Term Care（Canada）[39]，Endocrine Society Clinical Guideline（USA）[36]，Kidney Disease：Improving Global Outcomes[40]，Evidence-based Urology（EBUro）の一つである American Urological Association[41]，Society of Critical Care Medicine（SCCM）[42]，U. S. Preventive Services Task Force（USPSTF）[43]，Agency for Healthcare Research and Quality（AHRQ）[44]，あるいは非公式であるものの National Institutes of Health and Clinical Excellence（NICE）[4]，など約 25 の医療機関や団体が，GRADE システムを，オリジナルあるいは改変システムとして採用している（2007 年 9 月時点）。いずれにも共通していることは，GRADE システムの利便性や客観性，明瞭性などを評価したもので，あらゆる診療分野で利用され，急速な普及をみており，臨床現場での推奨度の評価や診療ガイドライン作成，予防医学の領域での適用も報告されている。

GRADE ワーキンググループではメーリングリストを使っての意見交換，問題抽出，他のシステムとの協調性，GRADE の用語整理をしつつ，さらに近々 BMJ への 5 シリーズとして，本システムの基本，診断研究，経済評価などの論文発表を予定している[26,27,45〜47]。

さらに同グループが開発した GRADE *profiler* と Cochrane レビューの RevMan 5 との統合による SoF 表[4,21,22]利用普及や，今後の Guyatt らによる GRADE の発表で，さらに GRADE システムの普及は加速されるだろう。

2　国内における GRADE システム普及の遅れ

一方，国内では，GRADE システムに関しては，2005 年 4 月に厚生労働科学研究業績として，当時より GRADE メンバーであった津谷・島村らの BMJ 誌の紹介や業績班発表がある[48〜50]。しかし，GRADE システムの根幹ともいうべき推奨度についての記載が 4 分類としたもので誤解を与えやすいものであった。実際，その後作成された日本の診療ガイドライン，たとえば「小児急性中耳炎診療ガイドライン」や「褥瘡局所治療ガイドライン」では[51,52]，GRADE を参考・参照したと記載されているが，GRADE システムを適切に利用したエビデンスの質の吟味・評価はなく，むしろ従来からの研究デザインを主体としたもので，国内の現状を考慮するとやむをえないものとも思われる。

患者側因子を中心としたガイドラインの必要性は，国内でも従来から，診療ガイドライン作成の手順（GLGL）ver4.3 2001（福井・丹後）[14]で強調されていたが，実際に作成された国内診療ガイドラインの多くは，アウトカムの重要度や，患者因子を反映したものではない。たとえば，アウトカムの重要度に関する検討はほとんどなく，患者参画での診療ガイドラインはわずかに 3 件と報告されている[53]。近々，国内の診療ガイドライン作成手順の改訂版が出る予定であるが，国際的な流れを考慮した，より単純・明瞭で，医療の最前線での利用性が高いものとして，統一性のあるシステムとの協調を考慮したものにすべきと思われる。

Ⅶ　GRADE システムの問題点

従来のガイドライン作りにおける推奨度評価

システムと比較して，GRADE システムは単純性，明瞭性などの面で，多くの利点が報告されている。しかし，注意すべきは，エビデンスの質評価における定量化に重点を置くことであり，単純なスコアリングによる判定にならないようにすべきであり，十分な検討が必要である。

国内において，本システムを採用・利用する上では，国内の多くの診療ガイドラインが作成基準も含め統一されておらず，GRADE 関連用語の理解など非常に多くの支障があり，また国内ガイドラインとの比較・変換という点でも容易ではない。また，今後も日本独自のグレーディングシステムを継続することは，国際会議あるいは海外論文との比較でギャップを生じることになり，非常に深刻な問題となるであろう。いずれにしても，エビデンスの質の評価や患者因子を考慮した GRADE システムの採用検討，コクラン・レビューの RevMan 5 対応 SoF table の理解のためにも，早急に本システムについての適切な理解が必須であると思われる。

ま と め

本内容は，国内診療ガイドライン作成などへの GRADE システム採用を強要するものではなく，本システムの適切な理解と普及，さらには SoF の理解・利用に役立つものとしての紹介を目的としたものである。本稿で説明できなかった部分，近々に公開される本システムの FAQ 和訳版，および今後登場する GRADE *profiler* の使用などを下記の URL で紹介していきたい。
http://homepage3.nifty.com/aihara/index.html
（なお，本論文の受稿後に，Cochrane Handbook ver. 5 ドラフト[22]が公開された）

文　　献

1）Atkins D, Best D, Briss PA, Eccles M, Falck-Ytter Y, Flottorp S, et al. GRADE Working Group. Grading quality of evidence and strength of recommendations. BMJ 2004;328:1490-4.
2）Oxman AD, Guyatt G, Schünemann H, on behalf of the GRADE Working Group. Grading the quality of evidence and the strength of recommendations. Cochrane Collaboration Methods Groups Newsletter 2004;8:6-7.
3）Kho M, Choong K, Friedberg J, Weeks J, Schunemann H. Application of GRADE in hematological malignancy：watchful waiting vs. immediate systemic treatment in indolent non-Hodgkin's lymphoma. *Cochrane collaboration*（*colloquia*, 2004）.
4）Schünemann H, Guyatt G. Cochrane Applicability and Recommendations Methods Group. Cochrane Collaboration Methods Groups Newsletter 2007; 11:18-9.
5）Institute of Medicine. Clinical Practice Guidelines：Directions for a new program. Washington DC：National Academy Press；1990.
6）Canadian Task Force on the Periodic Health Examination. The periodic health examination. CMAJ 1979;121:1193-254.
7）National Guideline Clearinghouse. NGC Browse — Organizations. Available：www.guidelines.gov/browse/browse.aspx
8）Center for evidence-based medicine. Levels of evidence and grades of recommendation. Oxford：The centre. http://www.cebm.net/index.aspx?o=1025（accessed 2007. Aug）.
9）Guyatt G, Rennie D. Users' Guide to the Medical Literature：A manual for Evidence-Based Clinical Practice. Chicago IL；AMA Press：2002.
10）Roman SH, Silberzweig SB, Siu AL. Grading the evidence for diabetes performance measures. Eff Clin Pract 2000;3:85-91.
11）Woloshin S. Arguing about grades. Eff Clin Pract 2000;3:94-5.
12）Guyatt G, Schünemann H, Cook D, Pauker S, Sinclair J, Bucher H, et al. Grades of recommendation for antithrombotic agents. Chest 2001;119:3-7S.
13）Harbour R, Miller J. A new system for grading recommendations in evidence based guidelines. BMJ 2001;323:334-6.
14）診療ガイドラインの作成の手順　ver. 4.3（福井・丹後 2001）　http://www.niph.go.jp/glgl-4.3rev.htm
15）Schünemann HJ, Best D, Vist G, Oxman AD, for the GRADE Working Group. Letters, numbers, symbols and words：how to communicate grades of evidence and recommendations. CMAJ 2003;169(7):677-80.
16）Upshur RE. Are all evidence-based practices alike? Problems in the ranking of evidence. CMAJ 2003: 169:672-3.
17）DISCERN：Health information on the Internet. Available：http://www.discern-genetics.org/index.php（accessed 2007 Sep.）.
18）Guyatt G, Baumann M, Pauker S, Halperin J,

Maurer J, Owens DK, et al. Addressing resource allocation issues in recommendations from clinical practice guideline panels：suggestions from an American College of Chest Physicians task force. Chest 2006;129:182–7.
19) Guyatt G, Gutterman D, Baumann MH, Addrizzo-Harris D, Hylek EM, Phillips B, et al. Grading strength of recommendations and quality of evidence in clinical guidelines：Report from an American College of Chest Physicians Task Force. Chest 2006;129:174–81.
20) Aihara M, Guyatt G, Falck-Ytter Y, Hama R. GRADE system：Going from evidence to recommendations. The Informed Prescriber 2007;22:91–102.
21) Vist G, Oxman AD, Glasziou P, Higgins J, Schünemann HJ. Summary of findings table in Cochrane reviews：a pilot study. Cochrane Collaboration Methods Groups Newsletter 2006;10:4–7.
22) Higgins J, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions (ver 5) (updated Nov. 2007) http://www.cochrane.org/resources/handbook/handbook5drafts.htm, http://www.cochrane.org/resources/handbook/Handbook5PresentingV6.pdf
23) Nonoyama ML, Brooks D, Lacasse Y, Guyatt GH, Goldstein RS. Oxygen therapy during exercise training in chronic obstructive pulmonary disease. Cochrane Database of Systematic Reviews 2007, Issue 2. Art. No.：CD005372. DOI：10.1002/14651858. CD005372. pub2.
24) Guyatt G, Vist G, Falck-Ytter Y, Kunz R, Magrini N, Schünemann H. An emerging consensus on grading recommendations? ACP J Club 2006;140:A11–2.
25) Schünemann HJ, Jaeschke R, Cook DJ, Bria WF, El-Solh AA, Ernst A, et al, on behalf of the ATS Documents Development and Implementation Committee. An official ATS statement：Grading the quality of evidence and strength of recommendations in ATS guidelines and recommendations. Am J Respir Crit Care Med 2006;174:605–14.
26) Guyatt G, Oxman AD, Kunz R, Falck-Ytter Y, Vist GE, Liberati A, et al. Going from evidence to recommendations. BMJ Series-3/5submitted.
27) Guyatt G, Oxman AD, Kunz R, Vist GE, Falck-Ytter Y, Schunemann HJ. What is 'quality of evidence' and why is it important to clinicians? BMJ Series-2/5 submitted.
28) Glenton C, Underland V, Kho M, Pennick V, Oxman AD. Summaries of findings, descriptions of interventions, and information about adverse effects would make reviews more informative. J Clin Epidemiol 2006;59:770–8.
29) Karanicolas PJ, Davies E, Kunz R, Briel M, Koka HP, Payne DM, et al. The Pylorus：Take it or leave it? Systematic review and meta-Analysis of pylorus-preserving versus standard Whipple pancreaticoduodenectomy for pancreatic or periampullary cancer. Ann Surg Oncol 2007;14:1825–34.
30) Naylor CD, Chen E, Strauss B. Measured enthusiasm：does the method of reporting trial results alter perceptions of therapeutic effectiveness? Ann Intern Med 1992;117:916–21.
31) Rossouw JE, et al (Writing Group for the Women's Health Initiative Investigators). Risks and benefits of estrogen plus progestin in healthy postmenopausal women：principal results From the Women's Health Initiative randomized controlled trial. JAMA 2002;288:321–33.
32) Higgins JP, Thompson SG, Deeks JJ, Altman DD. Measuring inconsistency in meta-analyses. BMJ 2003;327:557–60. doi：10.1136/bmj. 327.7414.557
33) Furukawa TA, Watanabe N, Omori IM, Montori VM, Guyatt GH. Association between unreported outcomes and effect size estimates in Cochrane meta-analyses. JAMA 2007;297:468–70.
34) Thompson DC, Rivara FP, Thompson R. Helmets for preventing head and facial injuries in bicyclists. Cochrane Database Syst Rev 2000;(2):CD001855.57
35) Cannegieter SC, Rosendaal FR, Briet E. Thromboembolic and bleeding complications in patients with mechanical heart valve prostheses. Circulation 1994; 89:635–41.
36) Wierman ME, Basson R, Davis SR, Khosla S, Miller KK, Rosner W, et al. Androgen therapy in women：an Endocrine Society Clinical Practice guideline. J Clin Endocrinol Metab 2006;91:3697–710.
37) UpToDate. http://www.uptodate.com/service/editorial_policy.asp
38) WHO. Rapid Advice Guidelines on pharmacological management of humans infected with avian influenza A (H5N1) virus. http://www.who.int/medicines/publications/WHO_PSM_PAR_2006.6.pdf
39) Ontario Health Technology Advisory Committee (OHTAC). Gastric electrical stimulation. Health technology policy assessment. OHTAC Recommendation 2006：http://www.health.gov.on.ca/english/providers/program/ohtac/tech/reviews/pdf/rev_ges_081806.pdf
40) Uhlig K, Macleod A, Craig J, Lau J, Levey AS, Levin A, et al. Grading evidence and recommendations for clinical practice guidelines in nephrology. A position statement from Kidney Disease：Improving Global

Outcomes（KDIGO）. Kidney International 2006；70：2058-65. http://www.nature.com/ki/journal/v70/n12/abs/5001875a.html
41）Dahm P, Kunz R, Schünemann H. Evidence-based clinical practice guidelines for prostate cancer：the need for a unified approach. Current Opinion in Urology 2007;17:200-7.
42）http://www.sccm.org/SCCM/Publications/Critical+Connections/Archives/August+2006/SSCguidelines_Aug06.htm
43）Guirguis-Blake J, Calonge N, Miller T, Siu A, Teutsch S, Whitlock E, et al. Current Processes of the U. S. Preventive Services Task Force：Refining Evidence-Based Recommendation Development. Ann Intern Med 2007;147:117-22. http://www.annals.org/cgi/reprint/147/2/117.pdf
44）AHRQ. Value of the Periodic Health Evaluation：Evidence Report/Technology Assessment, No. 136. 2007. http://www.ahrq.gov/downloads/pub/evidence/pdf/phe/phe.pdf
45）Guyatt G, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE：An emerging consensus on rating quality of evidence and strength of recommendation. Series-1/5 BMJ submitted.
46）Schünemann HJ, Oxman AD, Brozek J, Glasziou P, Jaeschke R, Williams J, et al. GRADEing the quality of evidence and strength of recommendations for diagnostic tests and strategies. BMJ Series-4/5 submitted.
47）Guyatt G, Oxman AD, Kunz R, Jaeschke R, Helfand M, Vist GE, et al. Grading recommendations：Incorporating considerations of resources use. BMJ Series-5/5 submitted.
48）津谷喜一郎, 辻香織. GRADE によるエビデンスの質とお勧め度の強さ.「根拠に基づく診療ガイドライン」の適切な作成・利用・普及に向けた基盤整備に関する研究：患者・医療消費者の参加推進に向けて（中山班, 平成 16 年度　総括・分担研究報告書）. p.6-9.
49）津谷喜一郎, 中山健夫, 島村治子. エビデンスの質とお勧め度のグレーディング. 薬理と治療2005;33:1241-54.
50）正木朋也, 津谷喜一郎. エビデンスに基づく医療（EBM）の系譜と方向性：保健医療評価に果たすコクラン共同計画の役割と未来. 日本評価研究2006;6:3-20.
51）小児急性中耳炎診療ガイドライン. 日本耳鼻科学会, 日本小児耳鼻咽頭科学会, 日本耳鼻咽頭科感染症研究会, 編集. 金原出版；2006.
52）日本褥瘡学会編. 科学的根拠に基づく褥瘡局所治療ガイドライン. 日本褥瘡学会. 昭林社出版；2005.
53）栗山真理子.「診療ガイドライン作成過程への患者・支援者参画のためのガイドライン」（PIPG）作成に至るまでの動機と経緯について. 厚生労働科学研究費補助金　（医療安全・医療技術評価総合研究事業),「根拠に基づく診療ガイドライン」の適切な作成・利用・普及に向けた基盤整備に関する研究：患者・医療消費者の参加推進に向けて（中山研究班平成 18 年報告書）. 2007. p.126-8.

GRADE System and SoF

—Going from Evidence to Recommendation (Ⅱ)—

Morio Aihara[1], Gordon Guyatt[2] and Yngve Falck-Ytter[3]

[1] *Aihara Clinic*
[2] *Department of Clinical Epidemiology and Biostatistics, HSC-2C12, McMaster University*
[3] *Case Western Reserve University, School of Medicine, VA Medical Center Cleveland*

The background：Clinical practice guidelines are increasingly popular, but the large number of systems for grading quality of evidence and strength of recommendations is confusing.

Purpose and significance：The GRADE (the Grading of Recommendations Assessment, Development and Evaluation) system of grading quality of evidence and strength of recommedations has been adopted by over 25 organizations including the Cochrane Collaboration, the World Health Organization (WHO), the National Institutes of Health and Clinical Excellence (NICE) in the UK, and UpToDate. In this article, we introduce the GRADE system.

1) The GRADE system illustrates how patient values and preferences play an important role when grading recommendations.

2) GRADE classifies patient-important outcomes as critical, important but not critical, and not important using a nine-point scale.

3) The strength recommendations, strong or weak, is determined by four factors (quality of the evidence, balance between desirable and undesirable effects, value/preference, and costs).

4) The quality of evidence (high, moderate, low, very low) is determined by evaluating the study design, and considering five factors that may reduce evidence quality (study limitations, inconsistency, indirectness, imprecision and reporting bias) and three that may increase quality including a large magnitude of effect and a dose-response gradient).

Recently, the useful tool for making the summary of findings (SoF) was developed by the GRADE working group and was integrated with the Review Manager (RevMan 5) of the Cochrane Collaboration. Guideline developers and clinicians can benefit from understanding the GRADE system using the SoF tables guide for their recommendations and clinical care.

＜2007年10月5日　受稿＞

表 7　用語集

用　語	意味・意義
absolute risk reduction（ARR）	絶対リスク減少率。対照群でのイベント発生率（CER）と治療群でのイベント発生率（EER）との差。すなわち，ARR＝CER－EER
association（関連）	厳密には関連の強さ（強固性）。特に観察研究で因果関係の強さを補強する因子。GRADE システムでは，グレードアップ評価因子の一つ。相対危険の程度により，強い関連性（RR＞2，＜0.5），非常に強い関連性（RR＞5，＜0.2）を判定する。large magnitude of effect とも記載される。
bias（バイアス）	研究の繰り返しによっても生じる推定値と真の値との差（系統的な誤差のこと）。対象者の選択やデータ収集・分析などに際して，方法が不適切な場合に生じ，結果を誤らせる重要な要因となる。ランダム誤差とは異なり，バイアスは真実からの一定方向への偏りをもたらす（すなわち，誤差が方向性を持つ）。治療効果についての研究では，バイアスの存在は，本来の作用が過大，あるいは過少評価になる。
binary data（二値データ）	二値データでは，多くの場合は，オッズ比を使って統合される。対数オッズは，差がなしを示す点から両側に距離の等しい対数スケールでプロットされる。オッズ比（対数）＝0 は両群間の差なしを意味する。dichotomous variable ともいう。
burdens（負担）	負担/重荷：患者あるいは介護人（家族など）が，推奨を固く守るようなことが嫌かもしれない要求で，回復には，より長期間かかるかもしれない治療の選択，より頻繁な検査を受けなければならない，などである。
confidence interval（CI：信頼区間）	真の値が存在すると期待される研究結果の範囲を示す。利益と害の両方の情報を提供する区間推定ともいう。
confounder（交絡因子）	関心あるアウトカムとも関連するために，関心のある変数の間の真の関係をゆがめる要因。confounding variable（交絡変数）ともいう。
continuous data（連続データ）	一定の間隔で並ぶ連続した数値で，血圧，体重，血球数など。数値変数 numeric variable，間隔変数 interval variable ともよばれる。
direction of confounder（交絡因子が影響する方向）	観察研究でのグレードアップ評価因子の一つ。結論に示された効果が，交絡因子のために真の効果よりも弱められていると考えられる場合。
directness（直接性）	グレードダウン因子で，EBM の PICO に相当するもので，研究の試験参加者，介入，アウトカム指標についての直接性である。
dose-response gradient（用量-反応勾配）	グレードアップの評価因子の一つで，用量－反応勾配があること（例，曝露量が多くなればなるほど，反応あるいは相対危険が大きくなる）。
downsides（マイナス面/短所）	害，リスク，負担（burdens），コストなど。
early stopping（早期中止）	RCT の中間解析で，有意な利益が得られたとして試験を早期に終了すること（また逆に害がある可能性が高くなったとして中断すること）がある。これは，計画した患者数より少ない患者数，短い観察期間となり，介入による利益がある試験の場合にはグレードダウンとなりうる。害のための早期終了はこの限りではない。
group sequential method（群逐次手法）	研究期間中にデータが徐々に蓄積されていくような実験や調査において，途中段階までに集まったデータをとりまとめて中間解析（Interim Analysis）を実施し，その結果をもとに実験や調査を早期に終了したり，研究デザインを修正したりする手法で，研究企画時に決定しておく必要がある。early stopping 参照。

表 7　用語集（つづき）

用　語	意味・意義
Guideline （ガイドライン）	GRADE システムによるガイドライン作成手順：(1) プロセスの確立，(2) 系統的レビュー，(3) 重要なアウトカムのエビデンス・プロフィール作成，(4) 各アウトカムについてのエビデンスの質，(5) アウトカムの相対的重要性，(6) 全体的なエビデンスの質，(7) 利益と害とのバランス，(8) 正味の利益とコストのバランス，(9) 推奨度決定，(10) 実施と評価。
Hazard ratio （ハザード比）	これまでに発生していない疾患がある短い時間間隔中に起こる速度を「罹患率（incidence rate）あるいは morbidity」という。同様に死亡がある短い時間間隔中に起こる速度を「死亡率（death rate）あるいは mortality」という。いずれもこれまで発生していない事象が発生する速度であることから，これらをまとめてハザードといい，これらを 2 群間で比をとったものがハザード比である。当然 HR が 1 ならば，群間に差はなく，1 点だけをみると，相対危険と同じである。
heterogeneity （異質性）	異質性は「不一致（inconsistency）」と同様のカテゴリーである。患者間あるいは研究間の相違を意味し，データ蓄積の信頼性の低さ，または不適切性につながる。(引用文献 25）では，$p<0.1$，あるいは I^2 値＞20％と規定している)。
imprecise （不精確）	信頼区間が広く，RR あるいは OR が 1 をはさんでおり，重要な害と重要な利益のどちらにも推定できる場合である。小さなサンプルサイズの研究など。
inconsistency （不一致）	複数の研究で推定効果・影響（の方向，程度）が一定していない（再現性が乏しい），あるいは異質性（heterogeneity）があること。
limitation （欠点）	ランダム化比較試験では，エビデンスの質のグレードダウンにつながるもので，割付の隠蔽やブラインド化あるいはフォローアップが不十分，両群間に重要な背景因子の偏りがある，利益があるとしての早期中止，ITT 解析でないものなど，結果の解釈に偏りをもたらすもの。risk of bias。
number needed to treat （NNT：治療効果発現必要症例数）	治療必要数，1 人の悪いアウトカムを避けるために治療が必要な患者の絶対数で，治療によるメリットに関する有用な指標の一つであるが，ベースラインリスクにより変化する。ARR の逆数である。害が生じる場合には，NNH という。
observational study （観察研究）	ある曝露要因の健康状態に対する関連を調べるための，意図的介入をしない（非実験的）方法による研究。コホート研究，症例対照研究，横断研究が代表的な観察研究の方法である。実験的研究/介入研究に対する用語。
outcome （アウトカム/結果）	GRADE システムでは，アウトカムの重要度を 3 段階で判定する：(1) 重大，(2) 重要だが重大でない，(3) 重要でない（GRADE *Profiler* では全体で，1〜9 ポイントをあて，最終的にはそれを三つのカテゴリーとする)。
outcome reporting bias （アウトカム報告バイアス）	報告バイアスの最も典型的な例。研究者（あるいは研究対象とした介入）にとって都合のよい結果（一部の病気や検査値の改善）は報告されやすく，より重要だが都合の悪い結果（死亡率の上昇や，重篤な害など）は報告されにくい。その最も著しい例が，研究報告バイアス（出版バイアス）である。
overall quality of evidence across outcome （アウトカムにわたっての全体的なエビデンスの質）	“推奨”を決定することがない（すべきでない）系統的レビュー作成者は，“複数のアウトカム”についての“全体的なエビデンスの質”は等級付けすることはなく，単に各々のアウトカムについてのエビデンスの質を等級付けする。ガイドラインパネルは，意思決定にとって本質的な，あらゆる重大な（critical）アウトカムにわたっての“全体的なエビデンスの質”を決定する必要がある。全アウトカムについての全体的なエビデンスの質を決定する際の原則は，

表 7　用語集（つづき）

用　語	意味・意義
overall quality of evidence across outcome（アウトカムにわたっての全体的なエビデンスの質）つづき	(1)“重大なアウトカム”についてのみ検討する。 (2) もし，複数の“重大なアウトカム”にわたってエビデンスが異なる場合； ・アウトカムが異なった方向ならば（利益と害の両方向），“最低のエビデンスの質”を採用する ・すべてのアウトカムが同じ方向ならば（利益，あるいは害），“最高のエビデンスの質”を採用する。
quality of evidence（エビデンスの質）	ある推定効果が正しいことにどの程度の確信（confidence）がもてるかを示すもの。GRADE システムでは，4 段階（high, moderate, low, very low）で評価する。
random error（ランダム誤差）	ある研究でなされた研究は，同等の患者から構成される集団で観察される様々な結果のほんの 1 例にすぎない。したがって，どんな標本集団での観察例の平均値も，母集団全体の真の値とは多少異なっている。偶然の誤差であり，バイアスと異なり，方向性はない。サンプルサイズの小さい研究ではバイアス（妥当性の低下）は生じないが，ランダム誤差のために誤った結果が生じる可能性が増加する。
recommendation（推奨）	利益が downsides（害・負担・コスト）を上回るか，その逆かに，どの程度確信できるかを示すもの。推奨度は 2 種類であり，その方向性も 2 種類である。推奨の強弱を判定するために四つの主要因子の考慮が必要である。
relative risk（RR：相対リスク）	risk ratio ともいう。治療群イベント発生率（EER）の対照群イベント発生率（CER）に対する比率。RR＝EER/CER
sparse（まばら）	一つの研究におけるアウトカム（結果）に関係した事象（イベント）の件数が少ない場合，その事象に関する情報が提供されない場合をいう。あるいは，ある結果に関する研究が 1 件しかない場合なども「まばら」というべきである。
standard mean differences（連続データでの標準化平均差）	測定尺度の単位が異なる場合，効果の大きさの要約推定値を求めるため，共通尺度として各研究での両群の平均値の差を標準偏差で割った値などを計算し，重み付けの和を計算するなどの統合方法をとる。
study design（研究デザイン）	基本的な研究デザインのことで，GRADE システムでは，ランダム化比較試験，観察研究および，その他の 3 種類に分類される。
study quality（研究の質）	エビデンスの質の評価の一つであったが，近年の報告では，欠点（limitation），risk of bias の有無とその程度によって評価される。
systemic review（系統的レビュー）	焦点をしぼった臨床疑問について，バイアスの可能性の低い方法を用いて答えようとする，研究の批判的評価。単なるサマリーではない。
trade-offs（トレードオフ）	利益大きくしようとするとそれに伴って害も大きくなり，害を小さくしようとすると，利益も小さくなるという関係。
validity（妥当性）	測定しようとした変数を正しく測定できている度合い，あるいは介入で達成しようとしたことが達成されている度合いをいう。
weighted mean difference（WMD：重み付け平均差）	アウトカムが二分的（死亡や心筋梗塞など）ではなく，連続的（症状スコアや身長など）な場合に用いられるエフェクト・サイズの尺度の一つ。研究対象の群間におけるアウトカムの平均差は，異なるサンプルサイズと精度を計算に入れて重み付けされる。WMD は絶対値であり，元のアウトカムの尺度の単位をもつ。