

## 0.3 GRADE システムに関する FAQ

GRADE ワーキンググループ作成の「よくある質問 (FAQ)」を、日本語翻訳版<sup>d</sup>にしたがって紹介する。この GRADE の FAQ は、2007 年に公開されたもので、やや古いですが、基本的なことを知るのに役立つだろう。

- Q 1 GRADE ワーキンググループとは何か
- Q 2 医学的なエビデンスをグレーディングすることがなぜ重要なのか
- Q 3 エビデンスと推奨をグレーディングすることは、何か新しいことなのか
- Q 4 世の中には、エビデンスと推奨のグレーディング・システムがたくさんある。なぜ、さらに別のものが必要なのか
- Q 5 過去にグレーディング・システムの開発に携わったことのある者は、GRADE ワーキンググループに参加してはならないか
- Q 6 GRADE とは、何の略か
- Q 7 エビデンスと推奨を系統的にグレーディングする長所は何か
- Q 8 「推奨の強さ (strength of recommendation)」とは何か
- Q 9 推奨の数が多いと、複雑ではないか
- Q 10 何が強い推奨とされるか
- Q 11 何が弱い推奨とされるのか
- Q 12 推奨の強さに影響する要因は何か
- Q 13 利益、リスク、負担、コストの大きさがどの程度確信をもてるかを決める要因は何か
- Q 14 基本的な研究デザインの違いとは、何のことか
- Q 15 エビデンスの質を決定するには、基本的な研究デザイン以外にどんな要因が重要か
- Q 16 GRADE では、エビデンスの質はどのように分類されるのか
- Q 17 エビデンスの質をグレーディングするというのは、ずいぶん抽象的に聞こえる。たとえば、「エビデンスの質が中 (moderate) である」とはどういう意味か
- Q 18 必要とされるすべてのエビデンスはどこで見つけるのか
- Q 19 ランダム化比較試験に欠陥 (flaw) があるときはどうか
- Q 20 ランダム化比較試験のどんな欠陥がエビデンスの質を下げるか
- Q 21 もし、ある研究では利益が証明され別の研究では利益が証明されない場合のように、複数の研究結果において一貫性がない場合はどうするか
- Q 22 ランダム化比較試験によるエビデンスが、自分が興味をもつ対象集団と類似してはいるものの厳密に同じ集団から得られたものでないときはどうするか
- Q 23 ランダム化比較試験において、試験の参加患者数が非常に少なくイベント数が非常に少ない場合には、どうするか

<sup>d</sup> GRADE に関する FAQ は、<http://www.gradeworkinggroup.org/FAQ/index.htm> で閲覧可能である。将来日本語で紹介される予定であるが、暫定的に <http://www.grade-jpn.com/grade-faq-test2.html> で紹介している。

- Q24 観察研究で、エビデンスの質が「中」あるいは「高」となるのはどのような場合か
- Q25 稀かもしれないが、観察研究からのすべてのバイアスが、見かけ上の治療効果を過小評価するよう働いていたらどうしたらよいか
- Q26 エビデンスの質がアウトカムごとに異なるときはどうしたらよいか
- Q27 なるほど。では、診断精度（精確さ）の研究はどう扱ったらよいか
- Q28 GRADE の方法についての詳しい情報はるか
- Q29 そのソフトウェア (GRADE profiler) の価格はいくらか
- Q30 GRADE について、もっと詳しく知りたい場合はどこをみたらよいか
- Q31 この FAQ がない疑問についてはどうすればよいか

### Q 1 GRADE ワーキンググループとは何か

GRADE ワーキンググループは、医療で用いられる既存のグレーディング・システムの問題点を改善したいと願う人々により、非公式の共同グループとして 2000 年に発足した。当グループの目的は、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を開発することである。

### Q 2 医学的なエビデンスをグレーディングすることがなぜ重要なのか

医学的なエビデンス、つまりエビデンスに基づく推奨はさまざまな質のものがある。エビデンスの情報源は、実験室での小さな研究や症例報告から、バイアスが最小となるように良くデザインされた大規模臨床研究まで幅広い。質の悪いエビデンスは患者が最も興味をもっていることではない推奨を導く可能性があることから、その推奨が強い（確信をもって推奨できる）のか、弱い（推奨に確信がない）のかを知っておくことが必須である。

### Q 3 エビデンスと推奨をグレーディングすることは、何か新しいことなのか

そうではない。グレーディングの考え方は 25 年以上前から使われている。

### Q 4 世の中には、エビデンスと推奨のグレーディング・システムがたくさんある。なぜ、さらに別のものが必要なのか

混乱を避けるために、単一システムが必要とされているからである。その単一システムは、他のシステムの短所を回避し、かつ長所を盛り込んだものであるべきだ。グレーディング・システムの中には、エビデンスの質を決定するにあたり、他の重要な要因を考慮することを明示せずに研究デザインのみに基づくものもある。また、システムによっては、複雑すぎるものもある。既存のグレーディング・システムの分析からは、これらやその他の短所に対しては、これまでいずれのシステムでも適切に対処されていない（表 3）。

### Q 5 過去にグレーディング・システムの開発に携わったことのある者は、GRADE ワーキンググループに参加してはならないか

いいえ、参加可能だ。実際に、広く使われているグレーディング・システムの開発者の多くが、GRADE の開発に積極的に関与している。

### Q 6 GRADE とは、何の略か

Grading of Recommendations Assessment, Development and Evaluation の頭文字をとって、GRADE と呼んでいる。

### Q 7 エビデンスと推奨を系統的にグレーディングする長所は何か

推奨の強さを系統的にグレーディングする方法により、バイアスを最小化し、専門家が作った医療用ガイドラインを解釈するのを助けることができる。実際に、ほとんどのガイドライン作成グループは、なんらかの系統的なグレーディングの考え方が必要だと考えている。

表 3 GRADE と他のシステムとの比較

要因 (Factors)	他のシステム (Other systems)	GRADE	GRADE システムの長所 (Advantages of GRADE system) *
定義 (Definitions)	エビデンスの質 (レベル) と推奨の強さについての定義が黙示的である。	定義が明示的である。	各グレードの意味やグレードを決定する際に考慮すべき事項が明確である。
判断 (Judgments)	どのアウトカムが重要か、各重要なアウトカムのエビデンスの質、利益と害のバランス、増分利益の価値に関する判断が黙示的である。	連続的 (sequential) な明示的判断。	各判断が明確であり、黙示的な判断プロセスにありがちなエラーやバイアスのリスクを減らすことができる。
エビデンスの質の主要構成要素 (Key components of quality of evidence)	重要なアウトカムそれぞれのエビデンスの質を決める重要な要素が考慮されていない。エビデンスの質はしばしば研究デザインのみで判断されている。	研究デザイン、研究の質、一貫性、エビデンスの直接性について、系統的かつ明示的に考慮して、エビデンスの質を判断している。	エビデンスの質を決める重要な要素が適切に考慮されるようになってきている。
エビデンスの質に影響する他の要因 (Other factors that can affect quality of evidence)	明示的に考慮していない。	不正確なデータもしくは少ないデータ、報告バイアス、関連の強さ、用量反応勾配のエビデンス、ありそうな交絡因子を明示的に考慮している。	その他の要因も考慮されるようになってきている。
全体的なエビデンスの質 (Overall quality of evidence)	利益に関するエビデンスの質に基づいて黙示的に判断されている。	意思決定に重大 (critical) な影響をもつアウトカムに関する、最も質の低いエビデンス (lowest quality of evidence) に基づいて、判断されている。	重大なアウトカムに関するエビデンスが欠如している場合でも、全体的なエビデンスの質 (overall quality of evidence) に対して誤った評価がされにくいように作られている。
アウトカムの相対的重要性 (Relative importance of outcomes)	黙示的に考慮されている。	重大な (critical) アウトカム、重大ではないが重要な (important) アウトカム、重要ではなく (unimportant) 無視 (ignore) できるアウトカムなのか、が明示的に判断されている。	全体的なエビデンスの質と推奨の強さをグレーディングする際に、各アウトカムが適切に考慮されるようになってきている。
健康上の利益と害のバランス (Balance between health benefits and harms)	明示的には考慮されていない。	重要な利益と害の得失、それらのエビデンスの質、特定の状況におけるエビデンスの適用、ベースラインリスクの確実性が明確に考慮されている。	利益と害についての判断をより明白で透明性が高いものにしていく。
健康上得られる増分利益はコストに見合うか (Whether incremental health benefits are worth the costs)	明示的に考慮されていない。	正味 (net) の健康上の利益について考慮された後に、明示的に考慮されている。	正味の健康上の利益の価値について、透明性の高い判断が下せるようになっていく。
エビデンスと結果の要約 (Summaries of evidence and findings)	提示の一貫性がない。	質評価と結果の要約を含む、GRADE エビデンス・プロファイル (evidence profile) に一貫している。	誰でも入手できる同じ情報に基づいてすべてのパネルメンバーが、判断する。
使用範囲 (Extent of use)	複数の機関で使われることはめったになく、経験的評価がされているとしてもほとんどない。	幅広い領域の機関が国際的に共同して作成および評価を行っている。	これまでの経験に基づいて、より合理的で、信頼でき、幅広く適応可能なシステムを目指していること。
	* これらの長所のいくつかを組み込むものもあるが、他のシステムのほとんどの取り組みは、これらのどの長所も含まない。		

### Q 8 「推奨の強さ (strength of recommendation)」とは何か

介入を実施すべき、あるいは、実施すべきでないという推奨は、利益と、リスク・負担 (burden) ・潜在的なコストとの間の得失 (trade-off) に基づくべきだ。もし利益がリスクや負担を上回るなら、専門家は臨床医が典型的な患者にその治療を提供するよう推奨する。利益とリスクや負担との間の得失に関連する不確実性 (uncertainty) により推奨の強さは決まる。

### Q 9 推奨度の数が多いと、複雑ではないか

その可能性がある。そのため、GRADE では、「強い推奨」、「弱い推奨」という2段階である。

### Q10 何が強い推奨とされるか

入手可能なエビデンスに基づき、利益がリスクや負担を上回る、または上回らない、という強い確信を臨床医がもつものを、強い推奨とする。

例 心筋梗塞後のアスピリン

アスピリンの短期間使用は、心筋梗塞後死亡の相対リスクを約 25%減少させる。アスピリンの有害イベントはわずかで極めて低コストである。その選択の意味を理解するならば、事実上すべての心筋梗塞患者はアスピリンを服用することを選ぶというのが、一般の人々の価値観と好みだろう。それゆえ、この状況では臨床医はアスピリン投与を強く推奨できる。

### Q11 何が弱い推奨とされるのか

入手可能なエビデンスに基づいて、臨床医が、利益とリスクや負担のバランスがよくとれている (finely balanced)、もしくは利益と害の大きさがかなり不確実であると確信する場合には、弱い推奨を提示しなくてはならない。加えて、臨床医は、臨床決断 (clinical decision making) において患者の価値観や好み (values and preferences) の重要性をますます意識するようになっていく。十分に情報を与えられた (fully informed) 患者たちが、価値観の幅の中で、さまざまな選択をとる傾向があるとき、ガイドラインパネルは弱い推奨を提供すべきである。

例 特発性深部静脈血栓症に対するワルファリン

特発性深部静脈血栓症の 40 歳男性患者が、用量を調節したワルファリンを 1 年間服用しているとする。もしこ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 の再発リスクは 1 年あたり約 10%ずつ減少するだろう。ワルファリン治療による避けられない負担としては、毎日ワルファリン 1 錠を服用すること、ビタミンK摂取量を一定にすること、血液検査による抗凝固能のモニタリング、小出血および大出血のリスク増加をかかえて生活することなどがある。DVT の再発を強く嫌う患者によっては、ワルファリン服用による不利益 (downsides) は DVT 再発予防効果に十分値すると思うかもしれない一方で、ワルファリン服用によって得られる利益は、リスクや不便さに値するものではないと考える人もいよう。

### Q12 推奨の強さに影響する要因は何か

推奨をグレーディングするときに考慮する必要のある要因は、たくさんある。1つの問題は、利益と害の最良推定値 (best estimate) への確信の度合いである。エビデンスの方法論的な質を評価するには、まず、そのエビデンスのもたらす利益と害の推定がどの程度確実なのかを考えなければならない。しか

し、推奨の強さに影響を与える要因は、この他にもたくさんある (表 4)。

表 4 推奨の強弱を決定するのに考慮すべき項目一覧

項目	例
考えられる利益と、考えられるリスク・不便さ・コストの推定値の不確実性 (Uncertainty)	多くの質の高いランダム化試験が喘息に対する吸入ステロイド療法に利益があることを示してきた一方で、症例集積研究しか気胸に対する胸膜癒着術の効用を検討してきていない。
治療が予防するアウトカムの重要性 (Importance)	血栓溶解療法で肺塞栓症死亡を予防することと、血栓溶解療法で深部静脈血栓症 (DVT) に対する静脈炎後症候群を予防すること。
治療効果の大きさ (Magnitude)	一過性脳虚血発作 (TIA) におけるクロピドグレル対アスピリンの脳卒中減少 [相対リスク減少 (RRR: 8.7%)] は、心房細動患者における抗凝固療法対プラセボの脳卒中減少 [相対リスク減少 (RRR: 68%)] よりも少ない。
治療効果の精確さ (Precision)	心房細動患者に対してアスピリンを投与した場合とプラセボを投与した場合とを比較した研究の結果の信頼区間は、一過性脳虚血発作 (TIA) 患者に対して脳卒中予防のためにアスピリンを投与した研究の結果の信頼区間と比べて広い。
治療に伴うリスク (Risks associated with therapy)	急性冠症候群に対する抗血小板療法としてアスピリンとクロピドグレルを併用した場合、アスピリンを単独投与するのと比べ、出血のリスクが高い。
治療の負担 (Burden)	ワルファリンの適量投与は、アスピリン治療に比べ負担が重い。ワルファリン治療は、抗凝固能とより継続的なビタミンK摂取量のモニタリングを必要とする。
標的イベントを起こすリスク (Risk of target event)	手術患者の中には術後の深部静脈血栓症や肺塞栓症のリスクが非常に低い人がいる一方で、深部静脈血栓症や肺塞栓症の発生率がかなり高い人もいる。
コスト (Cost)	一過性脳虚血発作 (TIA) 患者では、アスピリンに比べるとクロピドグレルはより高いコストがかかる。
さまざまな価値観 (Value)	若くて健康な人のほとんどは、延命することに高い価値をもつはずだ (そして、そのための苦痛を背負うことになる)。一方で、高齢者や衰弱した人は、延命の価値をどこに置くかはさまざまだろう (そして、そのために経験する苦痛への準備があるかも、人それぞれであろう)。

### Q13 利益、リスク、負担、コストの大きさがどの程度確信をもてるかを決める要因は何か

治療効果の有益性および有害性推定にどの程度確信をもてるかを決める際は、基本的な研究デザインと付加的方法論的要因 (methodological factor) が極めて重要になる。

### Q14 基本的な研究デザインの違いとは、何のことか

対象集団間の予後の違いや、盲検化のようなアウトカムの確認におけるバイアスを避けるための安全装置の欠如のため、一般的に、観察研究に基づくエビデンスは、実験的な研究デザインであるランダム化比較試験からのエビデンスよりも、かなり弱くなる、ということである。

**Q15 エビデンスの質を決定するには、基本的な研究デザイン以外にどんな要因が重要か**

近年、利益とリスクの推定が正しいという確信の程度に影響する多くの要因がより認識されるようになってきた。それらの要因とは、研究計画の質の低さ、バイアスの可能性が高い (high likelihood of bias) ことが示唆される入手可能なランダム化比較試験の実施、研究結果の非一貫性 (inconsistency)、エビデンスの非直接性 (indirectness)、まばらな (sparse) エビデンスといったことである<sup>e</sup>。

**Q16 GRADE では、エビデンスの質はどのように分類されるのか**

エビデンスのグレーディングの過程を経て、全体的なエビデンスの質 (overall quality of evidence) は、「高 high」、「中 moderate」、「低 low」、「非常に低 very low」に分類される。

**Q17 エビデンスの質をグレーディングするというのは、ずいぶん抽象的に聞こえる。たとえば、「エビデンスの質が中 (moderate) である」とはどういう意味か**

エビデンスの質をグレーディングする際は、以下の定義を用いる。

- ・「高」＝今後の研究によって効果推定値への確信性が変わる可能性は低い (very unlikely)。
- ・「中」＝今後の研究によって効果推定値への確信性に重要な影響がおよぶ可能性が高く (likely)、推定値が変わる可能性がある (may)。
- ・「低」＝今後の研究によって効果推定値への確信性に重要な影響がおよぶ可能性が非常に高く (very likely)、推定値が変わる可能性が高い (likely)。
- ・「非常に低」＝あらゆる効果推定値が不確実である (very uncertain)。

**Q18 必要とされるすべてのエビデンスはどこで見つけるのか**

理想的には、エビデンスの質をグレーディングする人たちは、必要とする人たちが検討している他の管理戦略選択肢の利益と害に関するエビデンスのシステマティック・レビューを入手できるようにすべきである。

**Q19 ランダム化比較試験に欠陥 (flaw) があるときはどうか**

重要な限界 (limitations) のあるランダム化試験については、エビデンスの質は「中」と分類する。深刻 (serious) な限界が多数あるランダム化比較試験については、エビデンスの質は「低」もしくは「非常に低」と分類する (表5)。

<sup>e</sup> 2010年2月時点では、エビデンスの質の評価を下げる5要因、評価アップの3要因がある (第1章, 1.5, 1.6)。

表 5 エビデンスのグレーディングの割り付け基準

エビデンスの種類 (Type of evidence)	<ul style="list-style-type: none"> <li>ランダム化試験 = 「高」</li> <li>観察研究 = 「低」</li> <li>その他のエビデンス = 「非常に低」</li> </ul>
グレードを下げるとき* (Decrease grade if)	<ul style="list-style-type: none"> <li>研究の質に (非常に) 深刻な限界 ((very) serious limitations) がある。</li> <li>重要な非一貫性 (inconsistency) がある。</li> <li>直接性 (directness) が、多少、もしくはかなり不確実である。</li> <li>データが不精確 (imprecise) もしくはまばら (sparse) である。</li> <li>報告バイアス (reporting bias) の可能性が高い。</li> </ul>
グレードを上げるとき* (Increase grade if)	<ul style="list-style-type: none"> <li>関連性 (association) を示唆する強いエビデンスがある：2 つ以上の観察研究による一貫性のあるエビデンスに基づいた相対リスクが &gt; 2 あるいは &lt; 0.5 で有意であり、これといった交絡因子がない (+1 段階)。</li> <li>関連性を示唆する非常に強いエビデンスがある：妥当性への大きな脅威がなく、直接的なエビデンスに基づいた相対リスクが &gt; 5 あるいは &lt; 0.2 で、有意である (+2 段階)。</li> <li>用量反応勾配 (dose response gradient) がある (+1 段階)。</li> <li>すべてのもっともらしい (plausible) 交絡因子 (confounders) が、真の効果をより弱めていると考えられるもの (+1 段階)。</li> </ul>
グレードの幅 (Range)	<ul style="list-style-type: none"> <li>エビデンスの質が 「高 high」</li> <li>エビデンスの質が 「中 moderate」</li> <li>エビデンスの質が 「低 low」</li> <li>エビデンスの質が 「非常に低 very low」</li> </ul>
* 各基準は、1 段階、もしくは非常に深刻なときは 2 段階、質を下げる。	

## Q20 ランダム化比較試験のどんな欠陥がエビデンスの質を下げるか

入手可能なランダム化比較試験に、治療効果のバイアスをもった評価をもたらす可能性の高い大きな欠陥があるとき、エビデンスへの確信の度合いが減る。このような方法論的な限界には、追跡率が非常に低いこと、割り付けの隠蔽が不適切であること、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盲検化がされていない研究、などがある。

例 ヘパリン誘発性血小板減少症 (heparin-induced thrombocytopenia: HIT) に対するダナバロイド治療

盲検化の欠如がグレーディングにどのような影響を与えるかについては、血栓合併ヘパリン誘発性血小板減少症 (HIT) に対するダナバロイドナトリウム治療の推奨をどうするかを考えてみればよい。HIT に対するダナバロイド治療についてのランダム化比較試験によるエビデンスは盲検化されていない研究によるもので、その試験のアウトカムはいつ血栓塞栓が消失したのかを臨床医が評価するという主観的なものだった。

## Q21 もし、ある研究では利益が証明され別の研究では利益が証明されない場合のように、複数の研究結果において一貫性がない場合はどうするか

複数のランダム化比較試験の間で治療効果推定値が大きく異なる場合 (結果に異質性 (heterogeneity) あるいは多様性があるとき)、研究者はその異質性を説明する理由を探す。たとえば、薬剤によっては、より重症の患者群あるいはより軽症の患者群において、相対効果が大きいことがある。異質性が存在するものの、研究者がその異質性についてのもっともらしい (plausible) 説明ができないとき、厳格なランダム化比較試験によるものとしても、推奨度はより弱くなる。

## 例 間欠性跛行の患者に対するベントキシフィリン

間欠性跛行の患者に対するベントキシフィリンのランダム化比較試験は、説明のつかない、矛盾する、結果を示した。原因不明の異質性を認めたことから、ACCP ガイドラインパネルはベントキシフィリンのエビデンスの質を「高」とはせず、「中」と判定した。

## Q22 ランダム化比較試験によるエビデンスが、自分が興味をもつ対象集団と類似してはいるものの厳密に同じ集団から得られたものでないときはどうするか

これは非直接的 (indirect) なエビデンスと考えるべきで、関連する集団への適用可能性は不確実であることから、エビデンスの強さのグレードを下げる必要がある。

## 例 外傷患者に対する段階的加圧ストッキング

段階的加圧ストッキングは、静脈血栓症のリスクをもつさまざまな集団に対して利益があることが証明されているが、利益があるかどうかは一度も直接外傷患者に対して試験されていない。ACCP のガイドラインパネルは、低分子ヘパリンの投与が禁忌である外傷患者に関する入手可能なランダム化比較試験を審査したが、他の集団からの一般化可能性に懸念があるため（つまり、エビデンスが非直接的である懸念があり）、エビデンスの質を「中」と判定した。もしエビデンスが直接性に懸念がなかったのならエビデンスの質を「高」と判定したはずである一方で、もし関連するランダム化比較試験が存在せず、最善のエビデンスが観察研究によるものだったのなら、エビデンスの質を「低」と判定したはずである。

## Q23 ランダム化比較試験において、試験の参加患者数が非常に少なくイベント数が非常に少ない場合には、どうするか

その場合もまた、エビデンスの質のグレードを下げる必要がある。こうした状況は、“まばらなデータ”と呼ばれることがある（表6）。

表6 不精確 (imprecise), あるいは、まばらな (sparse) データとは

データの不精確性やまばらなデータを定義するための経験的 (empirical) な基盤はない。定義するとすれば、以下の2つがある。

- データがまばら (sparse) である：研究結果にイベントや報告が数件しかなく、情報に乏しい場合。
- データが不精確 (imprecise) である：信頼区間の幅が、推定値が重要な利益あるいは害のどちらにも一致するほど広い場合。

これらの定義が異なれば、判断も異なってくる。定義を一致させるのは可能ではないかもしれないが、不精確またはまばらなデータによりエビデンスの質のグレードを下げるかどうかを検討するときのガイダンスとして下記を提案する。

- 1 件の研究しか存在しない場合、データが不精確、あるいはまばらであるとする閾値は下げるべきである。利益と害の両方にまたがる可能性がある幅広い信頼区間をもたらす、研究対象者数が小さい（あるいはイベント数が少ない）1 つの研究は、データが不精確、あるいはまばらであると考えべきである。
- 信頼区間の幅が広く、他のアウトカムとは無関係に、推定値が相反する推奨のどちらにも当てはまりうる場合、データが不精確、あるいはデータがまばらであると考えべきである。

## 例 脳静脈洞血栓症の患者に対するナドロパリン

例として、脳静脈洞血栓症の患者に対するナドロパリン（低分子ヘパリン）療法に取り組んだ、良くデザインされた厳格に実施されたランダム化比較試験について考えてみよう。ナドロパリンで治療された患者 30 人中 3 人に悪いアウトカムが生じ、同様に対照群の患者では 29 人中 6 人に生じた。研究者の解析では、悪いアウトカムの相対リスクは 52%減少したが、その結果は統計的に有意ではなかった。少ない患者数と少ないイベント数のため、ACCP ガイドラインパネルは脳静脈洞血栓症における抗凝固療法のエビデンスの質を、「高」ではなく、「中」と判定した。

## Q24 観察研究で、エビデンスの質が「中」あるいは「高」となるのはどのような場合か

観察研究のエビデンスの質は通常「低」のみであるものの、稀にエビデンスの質が「中」、あるいは「高」とさえ分類されることがある。たとえば、治療効果の大きさがとても大きく一貫した推定をもたらず稀な状況では、観察研究の結果に確信をもつかもれない。

## 例 心臓機械弁（人工弁）に対するワルファリン

これまで心臓機械弁の患者に対する抗凝固薬投与は、ランダム化比較試験でプラセボと比較されたことがない。しかし、観察研究によるエビデンスからは、抗凝固薬を投与しない場合の血栓塞栓イベントの発生率は、二尖式人工大動脈弁の患者で毎年 12.3%、他の 13 種類の弁ではさらに高いことが示唆され、抗凝固薬投与による相対リスク減少は 80%程度と推定された。観察研究では真の効果を過大評価しやすいことから、この観察研究という弱い研究デザインでは、全体的な利益説明することはできそうにない。そのため、ACCP ガイドラインパネルは、ランダム化比較試験は行われていないにもかかわらず、この観察研究のデータから、二尖式人工大動脈弁の患者に対する抗凝固薬投与の効果には強いエビデンスがあるとした。

表 5 エビデンスのグレーディングの割り付け基準（17 頁）参照

## Q25 稀かもしれないが、観察研究からのすべてのバイアスが、見かけ上の治療効果を過小評価するように働いていたらどうしたらよいか

つまり、実際の治療効果は研究データが示唆する治療効果よりも大きい可能性が非常に高いと考えられる状況である。実際には、この種のエビデンスの質のグレードを上げる必要がある。

## 例 私立の営利病院における死亡率

計 3,800 万人の患者を含む観察研究を用いた厳密なシステマティック・レビューにおいて、私立の営利病院と私立の非営利病院のケアを比較したとする。メタアナリシスでは、私立の営利病院の死亡率がより高いことが明らかとなった。レビューを行った研究者たちは、2 つのバイアスの可能性を主張した。1 つは、疾患の重症度による交絡が残っている可能性である。非営利病院の方が、どちらかという、営利病院の患者よりもより重症な可能性がある。こうした交絡が残っている限り、結果は非営利病院に不利な方向に偏ることになる。2 つ目に考えられるバイアスは、すばらしい民間医療保険保障をもつより多くの患者がよりよい医療資源をもつ病院へ行くことができ、それほど保障をもたない営利病院受診患者にも波及効果 (spill-over effect) をもたらず。営利病院は、非営利病院よりも、こういった補償範囲の広い保険をもつ患者を入院させやすいので、非営利病院にはまたしても不利な方向に偏ることになる。こうしたバイアスは、すべて、研究により証明される治療効果を減弱させる方向に振る舞うので、これらの観察研究から得られたエビデンスの質は、「低」ではなく、「中」と判定されるだろう。

## Q26 エビデンスの質がアウトカムごとに異なるときはどうしたらよいか

一般に、全体的なエビデンスの質は、意思決定を行う際に重大なすべてのアウトカムに関するエビデンスの質の中で「最も低いもの」とする。

例 サルコイドーシス患者に対するステロイド長期使用

ガイドラインパネルが、中等度から重度の症状とX線画像上の変化を認めるステージ2からステージ3のサルコイドーシス患者に対する経口ステロイドの長期使用について検討しているとする。ランダム化試験が、2年間にわたってX線所見、症状、肺活量検査に対するステロイドの影響を取り上げていたが、ステロイドの毒性については述べていなかった。もしガイドラインパネルが毒性を無視すれば、エビデンスの強さを「高」とするだろう。しかし、ガイドラインパネルが判断を下すために、ステロイドの毒性が極めて重要 (crucial) と考えるなら、ステロイドの影響に対する不確実性は増す。もしステロイドの毒性を推定するために観察研究を探すなら、毒性に関するエビデンスの質が低い可能性が高く、この低い評価が、最も適切なエビデンスの全体的な質となるだろう。代わりに、他の疾患に対するステロイドのランダム化試験を探しても、そのエビデンスの直接性の限界に直面するだろう。そこで、彼らは、ステロイドの毒性に関するエビデンス、および全体的なエビデンスの質を「中」とであると結論するだろう。

## Q27 なるほど。では、診断精度（精確さ）の研究はどう扱ったらよいか

精確な診断をすると、適切な治療を行うことができ健康アウトカムが改善する。また、診断が偽陽性であることによって余計な治療をしたことにより害が生じるといったことも、精確な診断を行うことにより減少する。こうした精確な診断を行うことによる健康アウトカムの改善や害の減少といった重要なアウトカムの代理指標として、診断検査の精度がある。しかし、診断検査のエビデンスが直接的かどうかは、その診断検査により、どのくらい正しく（真陽性または真陰性）、あるいはどのくらい誤って（偽陽性または偽陰性）分類され、その結果、どれほど重要な転帰がもたらされるかで判断するものである。GRADE ワーキンググループは、診断研究のエビデンスの質のグレーディングについて見識を深める文書を作成中<sup>f</sup>である。

例 急性尿路結石症の患者に対する非造影ヘリカル CT

急性尿路結石症疑いの診断において、非造影ヘリカル CTの方が静脈性腎盂造影よりも、偽陰性が少ないという、良くデザインされた複数の研究による一貫したエビデンスが存在する。しかし、このエビデンスが健康上重要な転帰をもたらすかどうかは、かなり不確実である。そのため、このエビデンスの質は推奨を行うにあたり「低」と考えられた。

## Q28 GRADEの方法についての詳しい情報はるか

GRADE ワーキンググループは、GRADEの使用を推進し、要約した表を作成するソフトウェア (GRADE profiler) を開発している<sup>g</sup>。

<sup>f</sup> Schunemann HJらにより、診断研究のGRADE利用論文が、BMJ 6編の論文の後も報告されている。GRADE: assessing the quality of evidence for diagnostic recommendations. (Evidence Based Med. 2008; 13: 162)<sup>32)</sup>

<sup>g</sup> 2008年に著者ら (Schunemann HJ, Brozek J, Oxman AD)の同意のもとでGRADE profiler (version3.2)のヘルプファイルに相当するGRADEハンドブックの日本語翻訳版を作成し公開した。ウェブサイト ([http://homepage3.nifty.com/aihara/grade\\_handbook3.2.html](http://homepage3.nifty.com/aihara/grade_handbook3.2.html)) からフリーでダウンロード利用可能であるが、2009年5月にアップデートされたversion 3.2.2とは内容が若干異なる。

**Q29 そのソフトウェア (GRADE profiler) の価格はいくらか**

そのソフトウェアは無料で以下から利用できる。

<http://www.cc-ims.net/revman/gradepro/gradepro>

**Q30 GRADE について、もっと詳しく知りたい場合はどこをみたらよいか**

GRADE を知るには、BMJ (<http://www.bmj.com/cgi/content/full/336/76501024>) 誌に掲載された入門編を読むとよいだろう。簡易化しつつもより詳細な解説は、現在作成中である<sup>h</sup>。

**Q31 この FAQ がない疑問についてはどうすればよいか**

<http://www.gradeworkinggroup.org/publications/index.htm> にたくさん情報がある。うまく情報を見つけれないときは、以下の電子メールで質問をお送りください：[mail@gradeworkinggroup.org](mailto:mail@gradeworkinggroup.org)

---

<sup>h</sup> 本 FAQ が公開された時点では、GRADE に関する BMJ 論文 (6 編, 2008 年)<sup>26-31)</sup> は報告されていなかった。