第7章　偶然性を考慮する

偶然性と大数の法則

治療効果に関するエビデンスの信頼性は、バイアス（偏り）の排除（そして、排除できなかったバイアスをどのように処理したか）にかかってきます。公正な検証にあるべきこうした特性を達成できなければ、研究結果をどのように分析したところで、時には死亡にさえつながる危険な問題が未解決のまま残ってしまいます（第 1 章および第 2 章を参照）。ただし、バイアスを減らす手段をとったとしても偶然性が働き、間違った方向に導かれてしまうこともあります。

コインを繰り返し投げたとき、表だけ、あるいは裏だけが 5 回以上「続いて出る」のもそれほど珍しくないと、皆さんご存知でしょう。そして、コインを投げる回数が増えるほど、最終的には表と裏が出る回数が同じような数になる可能性が高くなります。

2 つの治療法を比較する際、結果の違いは、単にこの偶然性が反映されている場合もあります。例えば治療 A を受けた患者の 40%が死亡したのに対し、治療 B を受けた同じような患者の 60%が死亡したとします。この 2 つの治療を、それぞれ 10 人の患者が受けたときは、表 1 に示したような結果が予想できます。2 つの治療における死亡者数の違いは、「リスク比」として示されます。この例でのリスク比は、0.67 になります。

こうした小さな数字をもとに、治療 A が治療 B より優れていると結論するのは適切でしょうか。おそらく違います。一方のグループ内の数人が、もう片方のグループより改善した背景には、偶然という理由が潜んでいる可能性があります。複数の他の小規模患者グループでもこの比較を繰り返したら、単なる偶然によって、それぞれのグループでの死亡者数が逆（6 人対 4 人）になったり、同数（5 人対 5 人）になったり、あるいはそれ以外の比率になるかも知れません。

	治療 A	治療 B	リスク比　（A:B=）
死亡者数	4	6	(4:6 =) 0.67
全体数	10	10	

表1．治療Aと治療Bの差について、この小規模研究から信頼できる推定が得られるでしょうか。

しかしそれぞれの治療を 100 人の患者が受け、各グループで先程と全く同じ比率の患者が死亡（40%と 60%）したらどうでしょうか（表 2）。表 1 に示した治療 A と治療 B を比較した際に生じた測定差（リスク比）は全く同じ（0.67）ですが、死亡者数が 40 人対 60 人の方が、4 人対 6 人に比べて明確な差であり、偶然性が反映された可能性も低いと思われます。治療法を比較する際に、偶然性に惑わされないようにするには、死亡、悪化、改善、同じ

状態のままという転帰をたどる十分な数の患者を含む研究をもとに、結論を導くことです。これはしばしば「大数の法則」と呼ばれます。

	治療 A	治療 B	リスク比 （A:B=）
死亡者数	40	60	(40:60 =) 0.67
全体数	100	100	

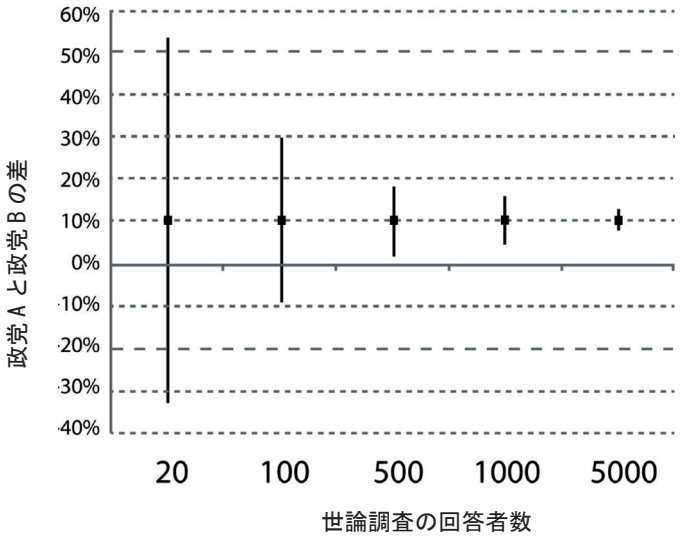
表2. 治療Aと治療Bの差について、この中規模研究から信頼できる推定が得られるでしょうか。

公正な検証でも偶然の可能性を検査する

治療法の公正な比較検証の結果を解釈する際にも、偶然性が働くことによって、私たちは 2 種類の間違いを犯してしまう可能性があります。2 つの治療法の転帰に本当は差がないのに、差があると誤って結論づけてしまう可能性、あるいは本当は差があるのに、間違って差がないと結論づけてしまう可能性です。観察できる対象の治療転帰の症例数が多いほど、このような間違いを犯してしまう可能性が減ります。

治療法の比較では、その治療を必要とする症状を過去に発症した、あるいはこれから発症する人全員を含めることはできないので、断定的にそれぞれの治療の「真の差」を見つけ出すのは不可能なのです。その代わりに臨床試験では、真の差にできる限り近い推定を出さなければなりません。

推定される差の信頼性は、しばしば「信頼区間（CI）」と表示されます。これは真の差の値が含まれる範囲を示しています。信頼区間という言葉を知らなくても、すでにほとんどの人はこの概念に馴染みがあることでしょう。例えば選挙運動期間に、政党 A が政党 B を 10%ポイントリードしているという世論調査があったとします。しかし調査報告には、しばしば政党間の差は最小で5ポイント、最大で15ポイントといった注釈がついています。この「信頼区間」は、政党間の真の差が、5%ポイントと 15%ポイントの間に位置している可能性が高いということを示しているのです。世論調査に回答した人の数が多いほど、結果についての不確定要素が小さくなるので、推定される差の信頼区間も狭まります。



政党Aと政党Bの差の95%信頼期間（CI）は、世論調査の回答者数が増えるほど狭くなる。

2つの政党支持者の割合に関して、推定される差を取り巻く不確定要素の度合いを評価できるように、2つの治療法を受けた後に悪化する患者と改善する患者の割合でも、推定される差に関わる不確定要素の度合いを評価することができるのです。そしてここでも、例えば心臓発作後の回復過程で2つの治療法を比べる場合に、治療転帰を観察できる患者の数が多いほど、推定される差の信頼区間は狭くなります。信頼区間は「狭いほど良い」のです。

通常、信頼区間は、その推定範囲内に真の値が入っていることに、どれだけ自信があるかも示されています。例えば「95%信頼区間」というのは、その真の値がこの信頼区間の中に入ると推定することに、95%自信があることを意味しています。これは100回のうち5回の割合（5%）で、「真」の値がその範囲外に位置する確率もあるということなのです。

治療間の「有意な差」とはどのような意味か

これは少し気をつけるべき質問です。というのも「有意な差」には、いくつかの意味があるからです。まず、患者にとって実際に意義のある重要な差という意味にも受け取れます。しかし研究文献の著者が「有意な差」があると表現するときは、しばしば「統計学的に有意」であるということを意味しています。そして「統計学的に有意な差」は、一般的な感覚での「有意義」にはあたらない場合もあります。治療間で、偶然である可能性が非常に低い差が「統計学的に有意な差」であって、実際にはあまり、あるいはほとんど意味がないこともあります。

1日1錠のアスピリンを摂取した何万人もの健康な男性と、アスピリンを摂取しなかった何万人もの健康な男性のランダム化比較試験に関する系統的レビューの例をみてみましょう。このレビューの結果、アスピリンを摂取した人の方が心臓発作を起こす確率が低く、

差は「統計学的に有意」であることがわかりました。つまりこの差が偶然性によるものだったと言える可能性は低いということです。ただし、それはこの差に実際意味があるとは限りません。健康な男性が心臓発作を起こす率がすでに非常に低いとすれば、薬を飲んでさらにその確率を下げるのは適切ではないかもしれません。特にアスピリンには、例えば出血などいくつかの副作用があり、時には死亡につながることもあります[1]。 系統的レビューから得られたエビデンスをもとにすると、もし 1,000 人の男性が 1 日 1 錠のアスピリンを 10 年間摂取して、うち 5 人がその期間に心臓発作を起こすのを回避できても、3 人に重篤な出血が起こると推計できるのです。

「統計学的に有意」とはどういう意味か

「正直なところ、注意を要する考え方だ。『統計学的に有意』とは、薬剤とプラセボの差や、2 つの集団に属する人の余命の差が、例えば、単に偶然によるものであるかどうか私たちに教えてくれるものである。（中略）『統計学的に有意』とは、単なる偶然だけで起きることはありえそうにないほど大きな差であることを意味する。

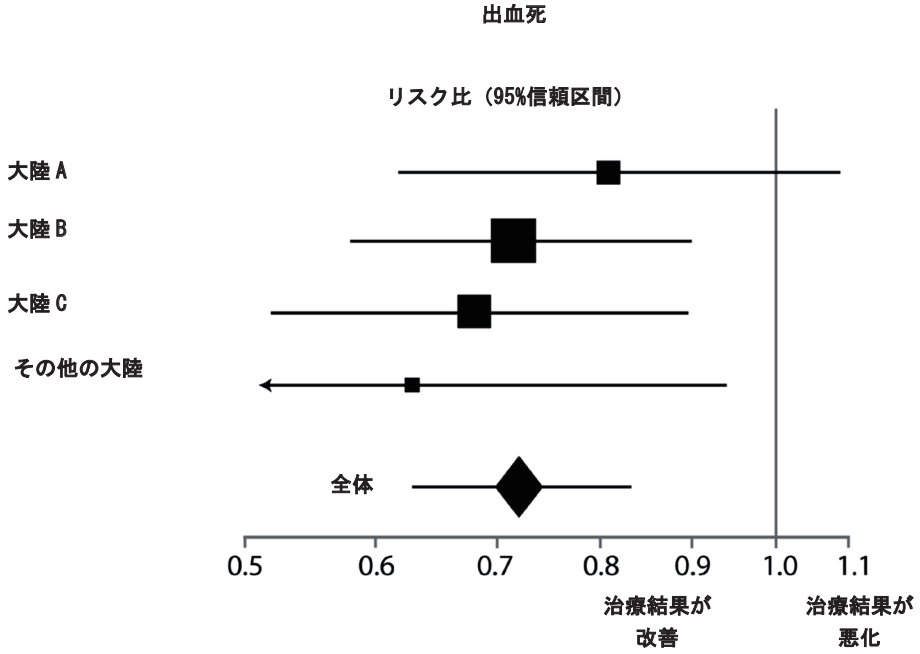
統計家は、この『ありえそうにない』という基準を使う。一般的には 5%水準（p=0.05 と書かれることもある）を有意とする。この場合、たとえ起きていることすべてが偶然であるとしても、それが起こる可能性は 20 分の 1 の割合未満なので、差が『有意』とされるのである」

Spiegelhalter D, quoted in:Making Sense of Statistics.2010.
www.senseaboutscience.org

治療の公正な検証のため十分な数を確保する

治療の比較試験では、1 つか 2 つの施設で試験を行うことで十分な数の参加者を確保することが可能な場合があります。しかし、治療が死亡といったまれな転帰に与える影響を評価するためには、通常、多くの国の多数の施設で患者を募集し、研究に参加してもらって、信頼できるエビデンスを得る必要があります。例えば 13 カ国から 1 万人の患者が参加したことにより、重篤な脳損傷がある患者へのステロイド薬の投与は、死亡を引き起こすことがわかりました。これは 30 年間続いてきた治療でした[2]。同じ研究チームによる別の公正な検証では、40 カ国から 2 万人の患者の参加を得て、トラネキサム酸と呼ばれる安価な薬剤が外傷後の出血死を減らすことが示されました[3]。これらの試験は、偶然性が働くことによる不確実性およびバイアスを減らすよう設計された非常に公正な検証であり、世界中のヘルスケアに大きく関連する質の高いエビデンスを提供しました。実際、英国医師会雑誌（BMJ）が実施した投票調査では、トラネキサム酸のランダム化比較試験が、2010 年の最重要試験に選ばれました。

下図は、受賞チームから提供されたデータに基づいたものです。偶然の働きで間違った結果に導かれるリスクを下げるためには、できる限り多くの情報をもとに、治療効果を推定することがいかに重要かを示しています。一番下のひし形のマークは、トラネキサム酸の臨床試験の全体の結果で、同薬が出血死をほぼ 30%（リスク比は 0.7 を若干上回る）下げたことを示しています。大陸 A の施設では、効果があまり顕著ではない（これは統計学的に有意ではなく、真の効果を過小評価している可能性が高い）という推定を示しています。また「その他の大陸」区分の施設では、より顕著な効果があった（過大評価の可能性が高い）という推定を示していますが、全体の結果がこの薬剤の効果について最も信頼できる推定です。



参加者全体および大陸ごとに示した重篤な出血がある外傷患者の死亡におけるトラネキサム酸の影響
(unpublished data from CRASH-2:Lancet 2010;376:23-32.)

同じように、国際的な試験で多くの施設から得られたデータを組み合わせ、「メタアナリシス」と呼ばれる方法で、類似の複数の試験結果を統計的に結びつけ（第 8 章も参照）、偶然性を低減することができます。メタアナリシスは長年にわたり統計家が開発してきた手法ですが、1970 年代から米国の社会科学者をたちが初めて利用してから、医療研究者の間でも頻繁に活用されるようになりました。20 世紀の終わりまでには、メタアナリシスが治療の正しい検査をする上で重要な要素として広く受け入れられるようになったのです。

例えば、未熟児の場合「血中酸素の値がいくつだと重大な障害が発生することなく最も生存の可能性が高くなるか」という 60 年間答えの出ていない問いに取り組むため、5 カ国で別々の財源および実施者による 5 つの試験が行われました。血中酸素レベルが高すぎると視力を失う可能性があり、低すぎると死亡あるいは脳性麻痺を起こす可能性があります。

こうした虚弱な新生児では、異なる酸素レベルによって生まれる差は小さく、それを検知するには多数の試験参加者が必要です。そのため 5 つの各研究チームの責任者たちは、1 つの研究結果から個別に導き出される推定よりも信頼度の高い推定を提供するために、それぞれの研究から得られたエビデンスを統合することに合意したのです。[4]

キーポイント
• 利用できるエビデンスの質と量における信頼性を検査する際には、「偶然性」を考慮する必要がある。