

# ソース

<https://www.sciencedirect.com/science/article/pii/S0091743522001761>

(<https://www.sciencedirect.com/science/article/pii/S0091743522001761>)

## 要約1

はい、承知いたしました。

まず、ご指定の通り、今後の説明において「**compatibility**」という用語は「**コンパチビリティ (相性の良さ)**」と翻訳します。この用語は、データと仮説またはモデルとの間の適合度合いを示す尺度として提案されています。

次に、ソース中のイベルメクチンに関する節（セクション2）についての詳細な要約を、段落ごとに説明します。

### イベルメクチンに関する節（セクション2）の段落ごとの詳細な要約:

- **段落:** パンデミック中の2020年、Covid-19の予防や治療にイベルメクチンを含めるべきかという問題が非常に物議を醸す話題となり、多くのランダム化比較試験が実施されました。残念ながら、これらの研究の中には質が疑わしいものや、不正データがあるように見えるものさえあり、そのようなケースでは統計的な提示に関する議論は無意味です。そのため、本稿では、そのような懸念が提起されなかった、軽症Covid-19の症状に対する低用量イベルメクチンに関する広く引用された試験（López-Medina et al., 2021）を取り上げ、その統計結果の適切な解釈に焦点を当てることにします。
- **段落:** その試験では、イベルメクチン群の症状寛解までの中央値時間は10日（四分位範囲 IQR, 9–13）であったのに対し、プラセボ群では12日（IQR, 9–13）であったと報告されました（症状寛解のハザード比 [率]、1.07 [95% 信頼区間 CI, 0.87から1.32]；ログランク検定によるP値 = 0.53）。21日目までに、イベルメクチン群の82%、プラセボ群の79%が症状が寛解しました。アウトカムはイベルメクチン群でわずかに優れているにすぎませんでした：寛解期間中央値は $(12 - 10)/12 = 17\%$ 短縮され、寛解率は7%高く、21日目での寛解率は $(82 - 79)/82 = 4\%$ 多かったという結果です。抄録は「5日間のイベルメクチン投与は、プラセボと比較して、症状寛解までの期間を有意に改善しなかった」と結論づけています。「有意に」の意味は曖昧であり、P値が0.05を超えたことのみを意味すると理解される可能性もありますが、著者らはほとんどの読者が、実際の効果が臨床的に重要でない、あるいは実質的な重要性がないことを意味すると受け取ると考えています。
- **段落:** このような読者にとって、その引用は完全に誤解を招く可能性があります。なぜなら、その試験における実際の影響はかなりの統計的不確実性を伴うからです。結果は、効果の欠如、存在、あるいは方向性について保証を提供するにはあまりに不正確です。特に、統計はイベルメクチンに効果があることを示すのに失敗している一方で、試験で使用されたイベルメクチンプロトコルが、もし寛解率が30%増加すれば「重要」と見なされる場合に、症状寛解までの期間を臨床的に重要な程度に改善しなかったことを示すのに失敗しています。代わりに、これらの結果は、一部の以前の試験で報告されたような劇的な効果をイベルメクチンプロトコルが持っていたという考えを反駁しています。
- **段落:** 著者とジャーナルは、イベルメクチンを誤って「奇跡の薬」として促進することに反対するために、効果を示すことの失敗を強調し、効果がないことを示すことの失敗に言及しなかった可能性があります。彼らの選択は、報告が完全性と正確さによって導かれるべきか、それともより行動的・政治的な目標によって導かれるべきかという問題として見られうるでしょう。この点を例証するかのよう、一般的な医療ニュースレターはこの試験を「軽症COVID-19でイベルメクチンは失望—コロンビアの試験は失敗」という見出しで取り上げ、イベルメクチン患者は「プラセボ群よりも良くなかった」と断言しました。このような主張は統計結果を誤って伝えており、ランダム化試験における「ランダム」性とその結果に残される不確実性（ただし、観察研究よりも測定可能な不確実性ではありますが）を考慮に入れていません。例えば、イベルメクチン対プラセボの寛解率の比率に対する95%「信頼」区間（CI）は

0.87から1.32まで広がっており、これはその区間内のすべての率の比率が $p > 0.05$ であることを意味します。

- **段落:** したがって、論文で採用された従来の5%の有意性カットオフを使用すると、結果は、試験で使われたイベルメクチン治療プロトコルによる症状寛解率が13%低下から32%上昇までのあらゆる可能性と最も**コンパチブル (相性が良い)**であるか、あるいはそれらの可能性に対してほとんど証拠を提供しないと言えます。統計を定性的な臨床用語に翻訳するには、何パーセントの改善が「臨床的に有意」と見なされるかについて合意する必要があります。もしそれが25%であり、この試験の報告方法を決定するための伝統的な0.05基準を受け入れた場合、結果は優柔不断であったと言えます。25%改善に対するP値は0.14であり、データと非常に**コンパチブル (相性が良い)**である一方、イベルメクチンによる30%の高い回復率や10%の低い回復率でさえ $p > 0.05$ であり、したがって0.05の慣習によればデータと合理的に**コンパチブル (相性が良い)**であるからです。しかし、それならば、「軽症Covid-19におけるイベルメクチンの決定的な回答を提供するのに試験は失敗した」といったより正確な見出しによって注目を最大限に集めることができるのでしょうか？そうではないと著者らは考えており、曖昧な結果を明確な「知見」に仕立て上げることで注目を集めることが、伝統的な統計的誤解の背後にある犯人の1つであると信じています。それにもかかわらず、50%高い回復率に対しては $p = 0.0015$ が得られます。したがって、その治療プロトコルに対するそれほど大きな効果に対して、試験はあまり**コンパチブル (相性が良い)**ではないとも言えます。したがって、もし50%を大きいと見なすが30%はそうでないと思なされるならば、「軽症Covid-19におけるイベルメクチンの大きな利益に対する証拠を試験は見出した」という見出しは弁護可能でしょう。
- **段落:** 統計的不確実性を無視する問題は何十年も嘆かれてきましたが、パンデミックはそれが医療研究と実践にもたらす害を加速させました。それは結果の誤報だけでなく、科学への不信感を悪化させることによっても行われました。それがそうする一つの方法は、衝突がないところに衝突があるかのような外観を作り出し、イベルメクチンの肯定的な報告が自動的に拒絶、無視、または却下される一方で、否定的な報告が高名なジャーナルに掲載され、医療ニュースで大々的に報じられるといった陰謀説を煽ることです。しかし、他の試験の中には、(Ahmed et al., 2021) のような「有意な」報告も見つかります。
- **段落:** この研究では、「5日間のイベルメクチン群におけるウイルス排除は、7日目と14日目にプラセボ群と比較して有意に早かった (ハザード比 (HR) 4.1、95% CI 1.1–14.7 ( $p = 0.03$ ) およびHR 2.7、95% CI 1.2–6.0 ( $p = 0.02$ ) ) 」と述べられています。この研究で測定されたアウトカムは、症状寛解ではなく検査室の終点 (ウイルス排除) であり、その違いや他の違いが、なぜこの研究が「肯定的」に見え、もう一つの研究が「否定的」に見えたかを説明するかもしれません。しかし、臨床的意義に関して言えば、衝突のいかなる主張も0.05のカットポイントによって生み出された幻想でしょう。このウイルス排除研究はLopez-Medinaの症状研究よりもはるかに不正確であり、したがって疑わしいものから莫大なものまで広がる効果と**コンパチブル (相性が良い)**であり、境界線の効果の重要な範囲内で症状研究と重複しています。

最後に、ソース全体の詳細な要約を提供します。

### ソース全体の詳細な要約:

- 医療および健康科学ジャーナルにおける統計分析はしばしば誤解を招き、間違っていることがよく知られています。何十年にもわたる改革努力にもかかわらず、「有意性」「信頼」といった有害な専門用語と、null仮説への偏った焦点により、統計の指導と報告は混乱しており、結果を分かりやすい言葉で説明しようとする試みは不十分です。
- このような状況の中で、有意性検定やP値に加え、あるいはそれらに代わる区間推定の導入は良い進展でした。しかし、これらの区間も同様の誤解に悩まされてきました。伝統的な統計手法を他の方法 (純粋尤度法やベイズ法など) に置き換える試みは、ほとんど影響を与えていません。
- そのため、著者らは、**P値や信頼区間を禁止したり置き換えたりするのではなく、伝統的な専門用語をより正確で控えめな日常言語のラベルに置き換えることを提案しています。**具体的には、**P値や区間推定を、データと仮説またはモデルとの間の「コンパチビリティ (相性の良さ)」の尺度として記述することを推奨しています。**このような記述は、観測データと**コンパチブル (相性の良さ)**な可能性の全範囲を強調します。

- 統計的誤解の持続、特に「統計的有意性」や「信頼区間」と現実との関係に関するものは驚くべきことです。何百年もの間、警告が繰り返されてきたにもかかわらず、多くの医療ジャーナルではP値が0.05を超える結果を「効果なし」と報告したり、null値を含む95%区間推定を効果がないかのように扱ったりする有害な慣行が続いています。これは、結果が実際には実際の効果量についてかなりの不確実性を残しているにもかかわらず行われています。
- 著者らは、研究者が「知識に害を及ぼさないことを第一に」というモットーを採用し、統計用語を、標準的な手続きが何を示し何を示さないのかをより良く反映するように修正するキャンペーンを行っていると述べています。これは、統計的手法そのものを禁止したり新しい方法を追加したりするのではなく、P値や区間推定を日常用語での「有意」かどうか、何を確信できるか、何が存在するかあるいは重要かどうかの判断基準として使用するような有害な慣行を終わらせることを目指しています。これらの統計は、分析データと分析手続きの仮定との間の**コンパチビリティ (相性の良さ)**のごく狭い側面しか示さないため、可能な効果を追求するか、治療法を助言するかについての決定にはるかに不十分であるとしています。
- 改革の簡単なアプローチは、「有意性レベル」としてのP値や「信頼区間」としての区間推定値のラベリングがどのように誤解を招いてきたかを理解することです。そして、それらの用語を、データとの仮説やモデルの**コンパチビリティ (相性の良さ)**という概念に基づいた記述に置き換えることです。これは、カットポイントを参照せずにP値を**コンパチビリティ (相性の良さ)**の尺度として提示することにつながります。また、区間推定値を**コンパチビリティ (相性の良さ)**区間として提示することにつながります。これは、結果に**自信**を植え付けるのではなく、最高の観測データでさえ幅広い可能性と非常に**コンパチブル (相性が良い)**であることを読者に思い出させるべきです。
- 区間は依然として任意のカットポイントに依存しているため、この二分法による誤解を減らすために、null仮説だけでなく代替仮説に対するP値を示すことを推奨しています。
- P値はデータと仮説のコンパチビリティ (相性の良さ) の尺度として解釈すべきです。**P値が1は完全な**コンパチビリティ (相性の良さ)**を、0は完全な非**コンパチビリティ (相性の良さ)**を表します。P値の尺度の問題を解決し、反証主義的な目標にも対応する一つの尺度が、P値の変換であるサンプライザル (surprisal) またはS値です。**S値はP値の負の対数 (底を2とする) であり、ビット単位で、データが仮説に対して提供する情報の量を示します。**
- イベルメクチンの事例を用いてこれらの概念が説明されています。特に、Lopez-Medinaの試験結果は、イベルメクチンに大きな効果があるという考えに対してはあまり**コンパチブル (相性が良い)**ではないことを示しています。
- 統計的方法は科学の進歩に貢献しましたが、「統計的有意性」と「信頼」は研究報告に広範な損害を与えてきました。これらは「エラー率」に基づくとされていますが、実際には理想化された状況でのみ当てはまり、研究における実際のエラーとはほとんど関連がありません。したがって、これらのラベリングを放棄するよう求める声が支持されています。
- 改革への呼びかけは伝統主義者によって反対されると予想されますが、著者らは伝統との決別が遅れていると考えています。約70年間、強制された統計的専門用語と慣習が研究報告の記述に著しい歪みをもたらしてきたという証拠があります。
- 適切な報告は、結果を偽って二分化するのではなく、推定値の不確実性を強調すべきであると長年議論されてきました。健康医学における単一の研究はしばしば暫定的な結論しか得られません。サンプルサイズの制限や理想からの逸脱（治療不遵守、離脱など）が不確実性を生みます。著者らは、不確実性と暫定性をより明確に認識させ、「矛盾する知見」という主張がしばしば二分的な宣言と「知見」を同一視することによって生じる単なる産物である理由を説明するのに役立つ改革を目指しています。
- 継続的な編集方針に見られるように、「有意性」のカットポイントを強制し、それを満たさない研究を「否定的」と宣言する抵抗が存在します。これは、過去の医療史における抵抗（細菌説否定など）と類似しています。研究コミュニティは、統計的実践も同様の抵抗に苦しんでいることを認識し、正確な平易な言葉による結果の報告のために積極的に闘うべき時であると結論づけています。

## 要約2

はい、承知いたしました。ソースの情報と会話履歴に基づき、ご質問にお答えします。「compatibility」は「コンパチビリティ (相性の良さ)」と翻訳します。

この論文は、健康科学および医学ジャーナルにおける統計分析の報告が頻繁に誤解を招き、あるいは間違っている現状について論じています。何十年にもわたる改革努力にも関わらず、この基本的な状況は変わっていません。特に、「有意性」「信頼」といったジャーゴンに固執し、帰無仮説に偏った焦点が当てられることで、結果の説明を単純化しようとする試みが損なわれています。

このような状況の中、統計的誤解は驚くほど持続しており、特に「統計的有意性」や「信頼区間」と現実の関係に関する誤報が *foremost* です。1900年代初頭には既に、「観察結果が『統計的に有意でない』からといって『効果が観察されなかった』と結論づけること」や、逆に「『統計的有意性』を重要な効果と同一視すること」に対する警告が見られました。これらの注意喚起は、その後の数百回の論評、編集論説、ジャーナルガイドラインで繰り返されてきました。それでもなお、主要な医学ジャーナルはP値が0.05を超える結果を「効果なし」と表現したり、95%区間推定がnull値を含む場合に効果がないかのように記述したりする記事を掲載し続けています。しかし実際には、これらの結果は実際の効果量についてかなりの不確実性を残しています。

著者らは、「知識に害を与えない」という医学の慣行にならったモットーを採用するならば、たとえそれが真実だと考えることと一致しても、研究者やジャーナリストが研究が示す内容について不当な主張をすること *discouragement* し、ジャーナルによって強制される証拠の歪曲に *vigorously* 抗議し反対することが義務であると述べています。そのため、著者らは標準的な手順が実際に行っていること、および行わないことをよりよく反映させるために統計用語を改訂するキャンペーンを行ってきました。これは、統計量を禁止したり新しい方法を追加したりするものではありません。むしろ、P値や区間推定を、普通用語で何が「有意」であるか、何について「信頼」できるか、何が存在するかしないか、何が重要かそうでないか、といったことの仲裁役として使用する **有害な慣行を終結**させることを目指しています。これらの統計量は、データと分析手順の仮定との間の **コンパチビリティの非常に狭い次元**しか示さないため、可能性のある効果を追求するか、治療を推奨するかについての決定には **遠く及びません**。

幸いなことに、このような状況の中で一つ肯定的進展があったのは、有意性検定とP値に加えて、あるいはそれに代わる **区間推定の導入**です。しかし、区間も同様の誤解に悩まされています。伝統的な統計を競合するもの（純粋尤度法やベイズ法など）に置き換えるよう求める試みも、ほとんど影響がありませんでした。したがって、P値や信頼区間を禁止または置き換えるのではなく、著者らは伝統的なジャーゴンを、これらの統計量を **データと仮説またはモデルの間のコンパチビリティ (相性の良さ)**の尺度として記述する、より正確で控えめな普通言語のラベルに置き換えることを提案しています。この「コンパチビリティ (相性の良さ)」という言葉は、統計モデリングの文献で長く使用されてきたものです。このような記述は、**観察データとコンパチブルな可能性の全範囲を強調**します。さらに、P値の単純な変換である **surprisal (驚き値)** または **S-value**は、データがそれらの可能性に対して供給する情報がどれだけ多いか少ないかという感覚を提供します。著者らは、Covid-19のイベルメクチン治療の試験という、非常に論争的となっているトピックからのいくつかの例を使用してこれらの改革を示しています。

統計改革は、例えばフリークエンティスト手法をベイズ手法に置き換えるような極端なものである必要はありません。しかし、ほとんどの読者が専門用語による記述をいかに真剣に受け止めるかを認識する必要があります。「有意性レベル」としてP値を、「信頼区間」として区間推定をラベリングすることが、前述のような混乱や誤解を助長してきたのです。したがって、改革の単純な第一歩は、「有意性」と「信頼」という紛らわしく誤解を招くジャーゴンを、データと仮説またはモデルの **コンパチビリティ (相性の良さ)**という **論理的に弱い概念**に基づく記述に置き換えることです。これは、P値を **カットポイントに言及せず**（特にほとんどの場合0.05のデフォルトを使用するのではなく）、**コンパチビリティ (相性の良さ)の尺度として提示**することにつながります。また、区間推定を **コンパチビリティ区間**として提示することにもつながります。結果に信頼を植え付けるのではなく、**区間は最良の観察データでさえ広範囲の可能性と高度にコンパチブル**であることを読者に思い出させるべきです。

区間は依然として包含のための任意のカットポイント（再び、ほとんどの場合  $p > 0.05$  のデフォルト）に依存しており、これが95%「信頼」の主張を生み出します。この二値化による誤解を減らすために、著者らは **P値を帰無仮説だけでなく、代替仮説に対しても示す**ことをさらに推奨しています。特に、臨床的に最小限重要と見なされるであろう代替案に対してこれを行うべきです。これは、公開されている信頼限界や一般的なソフトウェア出力を使用して容易に行うことができます。

Box 2では、P値を使用してさまざまな可能性との**コンパチビリティをゲージ**する方法、および二値の surprisalまたは**S-value（負のbase-2 log P値）**  $s = -\log_2(p)$ を使用して**それらの可能性に対する証拠をゲージ**する方法について、より正確な記述が提供されています。S-valueは、P値の尺度の問題に対処し（例えば、P値のスケールでは0.999と0.95の差が0.001と0.05の差と同じ距離に見えても、意味合いは全く異なる）、証拠をゲージするための尺度となります。その単位（ビット）は、単純なコイン投げ実験で対応するP値を得るために必要な裏が出ない回数として解釈できます。

## イベルメクチンに関する節（Section 2）の詳しい要約

イベルメクチンに関する節 (Section 2) では、統計的報告の課題を具体的に示すために、Covid-19治療薬としてのイベルメクチンの試験が例として取り上げられています。

- **段落8:** 2020年のパンデミックにおいて、Covid-19の予防や治療プロトコルにイベルメクチンを含めるかどうかという問題は、**非常に論争的**となりました。これにより、さまざまなプロトコルの多くのランダム化比較試験(RCT)が行われました。残念ながら、これらの研究のいくつかは**疑わしい質**であり、場合によっては**不正なデータさえ含まれている**ように見えました。このようなケースでは、それらの統計的提示についての議論は無意味です。そこで、著者らは、そのような懸念が提起されなかった、軽症Covid-19の症状に対する低用量イベルメクチンに関する**広く引用された試験（López-Medinaら、2021）**を検討し、その統計結果の**適切な解釈**に焦点を当てています。
- **段落9:** López-Medina et al. (2021) の報告によると、イベルメクチン群における症状回復までの中央値時間は10日（四分位範囲IQR 9-13）であったのに対し、プラセボ群では12日（IQR 9-13）でした。症状回復のハザード（率）比は1.07（95% CI 0.87から1.32）、対数ランク検定によるP値は0.53でした。21日目までに、イベルメクチン群の82%とプラセボ群の79%が症状が回復しました。したがって、結果はイベルメクチン群で**わずかに優れていたに過ぎません**：中央値回復時間の17%短縮（ $(12-10)/12$ ）、回復率の7%増加、21日目の回復の4%増加（ $(82-79)/82$ ）です。アブストラクトの結論は「イベルメクチンの5日間コースは、プラセボと比較して、症状回復までの時間を**有意に改善しなかった**」というものでした。この文章は「有意に」の意味について**曖昧**であり、単にP値が0.05を超えたことを意味すると理解されるかもしれませんが、著者らはほとんどの読者がこの引用句を**実際の効果が臨床的に取るに足らない、または実用的な重要性がない**という意味だと受け取るだろうと考えています。
- **段落10:** そのような読者にとって、引用句は**完全に誤解を招く可能性**があります。なぜなら、試験における実際の効果は**かなりの統計的不確実性**に左右されるからです。結果は効果の不在、存在、あるいは方向性でさえ**保証するには不正確すぎる**のです。特に、統計はイベルメクチンに効果があることを示すのに失敗していますが、同時に、試験で使用されたイベルメクチンプロトコルが、もし回復率が30%増加することが「重要」とされる場合、**臨床的に重要な程度まで回復時間を改善しなかったことを示すのに失敗**しています。代わりに、それらは一部の初期試験で報告されたような**劇的な効果**という概念を否定しています（Hill et al., 2022）。
- **段落11:** 著者やジャーナルは、イベルメクチンを奇跡の薬として誤って宣伝することに反対するために、効果を示せなかった点を強調し、効果がないことを示せなかった点に言及しなかったのかもしれませんが。彼らの選択は、報道が完全性と正確さによって導かれるべきか、それともより行動的・政治的な目標によって導かれるべきかという問題と見なせるかもしれません。この点を例証するように、人気のある医療ニュースレターは、この試験を「イベルメクチン、軽症COVID-19で期待外れ—コロンビアの試験は失敗」という見出しで扱い、「イベルメクチン患者はプラセボ群より良くなかった」と主張しました（Walker, 2021）。このような主張は、ランダム化比較試験における「ランダム」の要素が結果に残す不確実性を考慮に入れていないため、**統計結果を誤って伝えています**。例えば、イベルメクチン対プラセボの回復率の比に関する95%「信頼」区間は0.87から1.32まで伸びており、これは**その区間内のあらゆる率の比が $p > 0.05$ であることを意味**します。
- **段落12:** したがって、論文で採用されている従来の5%の有意水準カットオフを使用すると、Lopez-Medina試験の結果は、イベルメクチン治療プロトコルによる症状回復率が**13%低下から32%増加までのあらゆるものと最もコンパチブル**であるか、あるいはそれらに対する**証拠をほとんど提供しない**と言うことができます。統計を質的な臨床用語に翻訳するには、何パーセントの改善が「臨床的に有意」と見なされるかについて合意する必要があります。もしそれが25%であり、この試験を報告するための従来の0.05基準を受け入れた場合、結果は**決定的ではありません**。25%改善のP値は0.14であり、データと

**非常によくコンパチブル**です。一方、イベルメクチンの回復率が30%高くても10%低くても、 $p > 0.05$ であり、0.05の慣習によって**データと合理的にコンパチブル**です。しかし、それでは「試験は軽症Covid-19におけるイベルメクチンについて明確な回答を提供できなかった」のようなより正確な見出しが、最大限の注意を引くだろうか？著者らはそうは思わず、**曖昧な結果を明確な「知見」に仕立て上げることで注意を引くことが、「有意性」の硬い閾値を軸とした従来の統計的誤解の背後にある犯人の一つだと考えています**。それにもかかわらず、回復率が50%高いという仮説については $p = 0.0015$ を得ており、これは治療プロトコルについて**それほど大きな効果に対しても試験が非常にコンパチブルではない**と言える。したがって、もし50%を大きいとみなし、30%をそうでないとみなすならば、「試験は軽症Covid-19におけるイベルメクチンの大きな利益に反する証拠を発見した」という見出しは弁護可能であろう。

- **段落13:** 統計的不確実性を無視する問題は何十年も嘆かれてきましたが、パンデミックはその害を医療研究と実践に加速させています。結果の誤報だけでなく、科学への不信感を悪化させることによってです。これは、**対立がないところにまるで対立があるかのように見せかける**ことによって、イベルメクチンの肯定的な報告が自動的に拒否、無視、または却下され、否定的な報告が主要ジャーナルに掲載されて医療ニュースで大きく取り上げられるという陰謀論的な説明を助長します。しかし、他の試験の中には、Ahmed et al. (2021)のように「有意な」報告も見られます。
- **段落14:** Ahmed et al. (2021)の研究では、測定されたアウトカムは症状回復ではなく**ウイルス量の減少**という検査値アウトカムでした。この違い、または他の違いが、この研究が「肯定的」に見え、Lopez-Medina研究が「否定的」に見えた理由を説明するかもしれません。しかし、臨床的意義に関して言えば、いかなる対立の主張も0.05のカットポイントによって生み出された**幻想**であろう。このウイルス量減少の研究は、Lopez-Medinaの症状研究よりも**はるかに精度が低く**、したがって、疑わしいものから巨大な臨床的意義にわたる効果と**コンパチブル**であり、境界線の効果の重要な範囲内で症状研究と**重複しています**。

まとめると、明確な結果を提供するように設計された例外的な研究も存在し、有効で精度の高いデータを収集できればその目標を達成するかもしれません。しかし、**大多数の研究は、付随する見出しが示唆するような明確な結論を提供するには十分な信頼できる情報を提供しません**。たとえ研究の整合性が疑われていない場合でもです。正直な報告の精神に則り、そのような研究は**明確な結果を提供するのではなく、研究合成に利用可能な証拠のプールに貢献するものとして扱われるべき**です。したがって、著者らは、より現実的で控えめな目標を奨励する統計的プレゼンテーションの改善が必要であると考えており、その方法を**コンパチビリティ (相性の良さ)**の言葉を使って説明しています。

## 全体の詳しい要約

この論文は、健康科学と医学研究における統計結果の報告が、長年にわたり**誤解を招き、しばしば不正確**であった現状に警鐘を鳴らしています。特に、「統計的有意性」や「信頼区間」といった伝統的な統計ジャーゴンの使用が、データの解釈において混乱や誤報を引き起こしていることが指摘されています。P値が0.05を超えた結果を「効果なし」と断定したり、信頼区間にゼロや1を含む場合に効果がないかのように扱ったりする慣行は、実際の効果量に関する**かなりの不確実性**を無視しており、繰り返し警告されてきたにも関わらず依然として広く見られます。このような状況は、研究報告の質を損なうだけでなく、科学に対する不信感を悪化させる一因ともなっています。

このような問題に対処するため、論文は統計量そのもの（P値や区間推定）を廃止するのではなく、それらを説明する**用語を変更**することを提案しています。具体的には、「有意性」や「信頼」といった誤解を招きやすい言葉を、より正確で控えめな「**コンパチビリティ (相性の良さ)**」という概念に置き換えることを提唱しています。これは、P値を単なる合格/不合格のカットポイント（特に慣習的な0.05）ではなく、**データと仮説またはモデルの間のコンパチビリティの尺度**として扱うことを促します。同様に、区間推定は「信頼区間」ではなく「**コンパチビリティ区間**」と呼ぶことで、結果に対する過信を招くのではなく、**観察データが広範囲の可能性と高度にコンパチブルであることを読者に思い出させる役割を強調すべき**です。

さらに、著者らはP値の二値化による誤解を減らすために、**帰無仮説だけでなく、臨床的に重要となりうる代替仮説に対してもP値を示す**ことを推奨しています。これにより、データが特定の効果量とどの程度コンパチブルであるかをより詳細に検討できるようになります。P値に加えて、P値の変換である**S-value**



(surprisal または 驚き値) の使用も提案されています。S-valueは、**データが特定の仮説に対して提供する情報量をビット単位で示すこと**で、P値の尺度の問題（P値の数値的な差が直感的な証拠の差と一致しないこと）を克服し、証拠をより直感的に把握するのに役立ちます。

イベルメクチンのCovid-19治療薬としての試験が具体例として取り上げられ、統計的報告における従来の慣行の問題点が示されています。特に、広く引用されたLopez-Medina試験の結果が「有意差なし」と報告された際に、それがしばしば臨床的に効果がないという意味に誤って解釈されること、しかし実際には**結果が不確実すぎて、臨床的に重要な効果がないことを断定するには不十分**であったこと、そしてニュース報道が見出しで結果を誤って伝えたことが指摘されています。この事例は、異なる研究間で「矛盾する」ように見える結果も、従来の「有意性」カットポイントによって生じる**幻想**であり、実際には**データが不確実であり、広範囲の効果量とコンパチブルである可能性が高い**ことを示しています。

論文は、単一の研究がしばしば**暫定的な結論しか提供できない**という現実を強調し、そのような研究を「明確な知見」としてではなく、**エビデンスのプールへの貢献**として扱うべきだと主張しています。統計的伝統は、正確な報告を妨げる抵抗勢力となっており、研究コミュニティは、ジャーナルや査読者が強制する有害な慣行に異議を唱え、**正確な平易な言葉による報告**のために積極的に戦うべきだと結論付けています。用語を「コンパチビリティ (相性の良さ)」などに変更することは、このような正確で責任ある報告を促進するための重要な一歩となります。

## 研究結果の広報に関する疑問への回答

イベルメクチンのCovid-19への効き目に関する研究のような事柄の広報では、正確さを重んじるか、それとも社会的な影響への配慮の側を優先するべきかという難しい問題に悩まされることは、ご指摘の通りです。著者らは、Lopez-Medina試験の報告において、ジャーナルや著者がイベルメクチンの誤った宣伝に反対するために、正確さよりも政治的目標を優先した可能性に触れています。しかし、結果的にどうするにせよ、従来の統計的有意性の「有意」や信頼区間の「信頼」という言葉の使用は誤解を増やすことは、ソースで繰り返し指摘されています。

これらの言葉を「**コンパチビリティ (相性の良さ)**」という言葉で置き換えたとき、どのように適切に研究結果の広報を行うことができるようになるでしょうか？

- 1. 不確実性の正直な表現:** 「コンパチビリティ (相性の良さ)」という用語は、統計的結論がデータと特定の仮説との間の**両立可能性の度合い**を示すものであることを明確に伝えます。これは、「有意である」や「信頼できる」といった言葉がしばしば示唆するような**断定的な結論ではない**ことを示唆します。研究結果が不確実である場合（Lopez-Medina試験のように）、単に「効果なし」と報告するのではなく、「データは効果の存在を強く支持しないが、**広範囲の効果量（例：13%減少から32%増加まで）とコンパチブル (相性が良い)**ままである」といった、**データが実際に示唆する不確実性の範囲**をより正直に伝えることが可能になります。
- 2. 二値化の回避:** P値を特定のカットポイント（特に0.05）に基づいて結果を「有意」または「有意でない」と二値的に分類することは、多くの誤解の源です。これは、「有意でない」結果を「効果なし」と誤って解釈することにつながります。**P値をコンパチビリティの尺度としてカットポイントなしで提示**し、さらに**複数の仮説（特に臨床的に重要な代替案）に対するP値を示す**ことで、データがどの効果量とどの程度コンパチブルであるかという、よりニュアンスのある理解を促進できます。S-valueは、特定の仮説に対してデータが提供する証拠の量をより直感的に示す補助的なツールとなります。
- 3. 誤った対立の解消:** 異なる研究の結果が「矛盾している」ように見える場合でも、しばしばそれは従来の「有意性」の閾値によって生じる**幻想**です。実際には、両方の研究の結果が、広範囲の効果量とコンパチブルであり、特に効果が小さいまたは中程度の範囲で重複している可能性があります。結果を「コンパチビリティ (相性の良さ)」の観点から報告することで、これらの見かけ上の「対立」は、単に**データが不確実であり、さまざまな効果量とコンパチブルである**という事実の異なる側面を示しているに過ぎないことを説明しやすくなります。
- 4. 研究の役割の再認識:** 「コンパチビリティ (相性の良さ)」という言葉は、単一の研究がデータとある仮説の**両立可能性**という狭い側面しか示せないことを強調します。これは、**単一の研究が明確な結論を提**

**供することは稀である**という現実を反映しており、研究結果を最終的な「知見」としてではなく、**エビデンスの累積に貢献する一つの要素**として扱うべきだという考え方を強化します。

したがって、「有意性」や「信頼」といった言葉を「コンパチビリティ (相性の良さ)」に置き換えることは、研究結果の広報を、過度に断定的な主張や誤解を招く二値的判断から脱却させ、**データが示す不確実性を正直に伝え、データと仮説の間の微妙な関係性をより正確に表現**することを可能にします。これは、科学的なコミュニケーションの透明性と正確性を向上させ、最終的には科学への信頼を再構築する一助となる可能性があります。