

論文とETIC 10

AIを使うと 判断は良くなるのか

平均点・比較相手・重大な見落としから読む人とAIの協働



Daiki Morimoto / Unknown Lab

ORCID: 0009-0009-6444-750X

Version 1.0 / 公開日 2026-09-25

DOI: 10.5281/zenodo.22948356

未査読 / 専門職としての著者確認済み

独立専門レビュー未実施 / AI支援執筆

CC BY 4.0

30秒で読む

AIを使った人の成績が、使わなかった人より高い。この結果には価値がある。しかし、それだけで「人とAIを組み合わせることが最良の方法だ」とは言えない。AI単独との比較、重要な誤り、所要時間まで確かめる必要がある。

106実験を統合したメタ分析では、人とAIの組合せは平均して人単独を上回った一方、人とAIのうち成績のよい方を基準にすると下回った。[1] 医師92人の無作為化比較試験では、GPT-4を使える群の症例課題の得点は高かったが、回答時間も長くなった。AI単独との差は明確ではなく、実際の患者の健康状態は測っていない。[2]

本稿が考えるのは、AIの利用を肯定するか否定するかではない。「良い回答が増えた」「重大な失敗が減った」「人の判断が良くなった」を分け、どの比較で何を確かめるべきかという問いである。

よくできた提案書の、何が良くなったのか

次の場面は、研究参加者の実話ではなく、論点を説明するための仮想例である。

会議に、AIを使って作られた提案書が出てくる。検討項目が増え、代替案も書かれ、文章は読みやすい。以前の短いメモより、完成度が高く見える。ところが、実行する人の時間が足りないことや、本人が望まない選択肢であることは、依然として抜けている。書いた人が重要な条件を理解したかどうか、文書だけでは分からない。

この経験は、医療に限らず、教育、事業計画、設計、行政などにも想像しやすいだろう。ただし、似た経験を思い浮かべられることは、同じ頻度で問題が起こる証拠ではない。そこで一般論を重ねるのではなく、人とAIの協働を実験で測った研究へ戻る。

評価には少なくとも三つの問いがある。第一に、AIは人の仕事を助けたか。第二に、人とAIを組み合わせることで、どちらか単独より良い成果が得られたか。第三に、その成果は、本人にとって重要な結果につながったか。この三つは関連するが、同じではない。

例えば、AIの案を採用して文書の得点が上がっても、第一の問いに答えたにすぎない場合がある。人が危険な提案を見抜けたなら、その役割は平均点に表れなくても重要かもしれない。反対に、読む負担だけが増えたなら、文章の充実を業務全体の改善と呼ぶことはできない。何を成功と数えるかを先に決める必要がある。

二つの研究を、なぜ一緒に読むのか

一般的な人とAIの協働を調べたVaccaroらの研究は、比較相手によって結論が変わることを示す。[1] 一方、Gohらの研究は、治療や検査の進め方を考える医師の課題で、その問題を具体的に検討できる。[2] 前者は広い見取り図、後者は一つの条件下での比較である。

後者を前者の単純な再現試験とは扱わない。メタ分析の文献収集期間は2023年6月までであり、Gohらの試験はその後に行われている。対象のAI、課題、採点方法も異なる。二つの効果を平均して、新しい総合効果を算出することもしない。

研究	対象・期間	この論評での役割
Vaccaroら、2024年	2020年1月1日から2023年6月30日までの査読済み研究。 74論文、106実験、370効果量	人単独との比較と、単独で優れた方との比較を分ける。
Gohら、2025年	米国の医師92人。論文記載の実施期間は2023年11月30日から2024年4月21日	複雑な症例課題の得点、時間、AI単独、害の評価を具体的に読む。

表1。原著のMethods、Resultsと補足資料から作成。[1,2] 370は参加人数ではなく、複数の条件・評価から得られた効果量の数である。Gohらの実施日には登録情報との差があり、後で区別して記す。

106実験が示した、比較相手の重要性

対象と選定過程

Vaccaroらは2023年7月に、ACM Digital Library、Web of Science、Association for Information Systems eLibraryを検索し、引用文献と後続引用も調べた。人を対象とする実験で、人単独、AI単独、人とAIの組合せの三条件の成績が分かり、効果量を計算できる研究を選んだ。本文では5,126件から74論文を採用したと報告している。[1]

同じ論文に複数の実験があり、同じ実験に複数の評価がある。この依存関係を無視して370件を独立した試験とみなすと、証拠の量を大きく見せてしまう。著者らは実験内の効果量のまとまりを扱う三層のメタ分析と、推定の不確かさを扱う頑健な分散推定を用いた。異なる尺度の成績は、Hedgesのgという標準化した差へ変換している。

gは、点数や正答率の差をばらつきに対して表した指標であり、百分率ではない。gが0.64だから正答率が64%増えた、という読み方はできない。また、異なる仕事の尺度を同じ形に直しても、その仕事が生活にとって同じ重要性を持つようになるわけではない。

「人を助ける」と「組合せが最良」は違う

著者らは、人とAIの組合せが人単独を上回ることをhuman augmentationと呼ぶ。ここでは「人への支援効果」と表す。human-AI synergyは、人単独とAI単独の両方を上回ることを指す。ここでは「組合せによる追加的な優位」とする。[1]

この定義は厳しいが、重要である。人よりAIが高得点を取る課題で、AIの答えを人がそのまま採用しても、人への支援効果は生じ得る。しかし、人を加えたことによる追加価値はまだ示されない。逆に、人とAIの組合せがAI単独に平均点で及ばなくても、人単独より改善したという事実が消えるわけではない。

比較・課題	Hedgesのgと95%信頼区間	読める範囲
人とAI 対 人単独	0.64 [0.53, 0.74]	平均して、人単独より成績が高い。
人とAI 対 単独で優れた方	-0.23 [-0.39, -0.07]	平均して、単独で優れた方には及ばない。
決定課題で、単独で優れた方と比較	-0.27 [-0.44, -0.10]	選択・判断課題の平均では追加的な優位を認めない。
創作課題で、単独で優れた方と比較	0.19 [-0.09, 0.48]	推定値は正だが、区間は0を含む。優位を確定できない。

表2。全体は370効果量、決定課題336、創作課題34。本文と Supplementary Tables S3・ S7に基づく。[1] 決定課題と創作課題の差は示されたが、「創作なら必ず相乗効果が出る」という結果ではない。

ここでいう「下回る」は、人とAIのうち成績がよかった方との比較である。「人と組むと、人単独よりもAI単独よりも悪くなる」と言い換えると、基準が変わってしまう。全体の結果は、まさに人単独より改善しながら、より厳しい基準には届かないというものだった。

平均値を、あらゆるAIの結論にしない

研究間の違いは大きい。ばらつきの程度を示すI²は、組合せによる追加的な優位で97.7%、人への支援効果で93.8%だった。これは「約98%の研究が失敗した」という意味ではなく、観察された効果の違いの多くが、標本誤差だけでは説明できないことを示す。[1]

著者らは課題の種類、人とAIの元の成績、説明の有無などを比較している。しかし、研究同士の特徴の違いは、無作為に割り付けられた条件とは限らない。人がもともと得意な課題で組合せの成績が良いからといって、訓練で人の能力を上げれば同じ効果が生じるとまでは言えない。課題の難しさやAIの使い方も一緒に違っている可能性がある。

また、研究の大部分は人が最終判断をする構成であり、仕事の一部を事前に人とAIへ分けた実験は3件、4効果量に限られた。分業は有望な仮説だが、この少数の結果だけで有効な標準手順を確定することはできない。説明や

確信度の表示について明確な差がなかったことも、あらゆる場面で説明が不要だという意味ではない。

収集期間は2023年半ばまでである。生成AI以外の方式や、AIの応答を実験的に模擬した条件も含まれる。2026年の特定のモデルの性能表として読むべきではない。むしろ、AIが更新されても残る「比較条件を揃えなければ追加価値は分からない」という測定上の問題を取り出すことに意味がある。

医師92人の試験では、何が改善したのか

対象、介入、比較条件

Gohらは、Stanford、Beth Israel Deaconess Medical Center、University of Virginiaのメールリストを通じて、内科、救急、家庭医療を専門とする医師を募集した。参加者は92人で、指導医等のattendingが67人、研修医が25人だった。一般の医師全体を無作為抽出した標本ではなく、参加を希望した人の集団である。[2]

医師を46人ずつ二群に無作為に割り付けた。一群はGPT-4をChatGPT Plusから利用でき、通常の資料も使えた。対照群は検索や診療情報など通常の資料を使えたが、LLMは使えなかった。AIなしとは、何の支援もないという意味ではない。

課題は、匿名化した実際の診療経験をもとに専門家が作った5種類の症例だった。情報が段階的に示され、その時点で治療や検査の進め方を回答する。新しい情報を読んだ後に前の答えを書き直すことはできない。遠隔または会場で監督者のもと、1時間のセッションで実施された。すべてを急いで終えるより回答の質を優先するよう求められていた。

条件	実施・採点された単位	比較上の注意
医師 + GPT-4	46人、178症例回答	AIの使い方は基本的な操作説明と技術支援。高度な入力訓練の効果を試したのではない。
医師 + 通常資料	46人、197症例回答	検索等は利用可能。LLMだけが利用できない。
GPT-4単独	5症例を各5回、25出力	研究チームが作った入力で生成。医師を割り付けた第三の群ではない。

表3。人数と症例回答数を分けた。[2] 症例は5種類であり、375種類の異なる患者を扱ったわけではない。単独AIの反復出力も25人の独立した参加者には相当しない。

本文のResultsには支援群176、対照群199との記述があるが、Table 2、図の説明、公開CSVは178と197で一致する。Unknown LabでCSVの400行を再集計し、上表の件数と各群の平均得点を確認した。したがって本稿は表・図・CSVの値を採用する。これは記述統計の照合であり、原著の混合効果モデルの独立再解析ではない。

登録情報にも日付の違いがある。ClinicalTrials.govは実際の開始日を2023年12月28日、完了日を2024年4月19日とする一方、本文の期間は前掲の通りである。初回提出は2023年12月20日、初回公開は2024年1月17日だった。[3] 本稿では両方を区別し、「開始前にすべての登録内容が公開されていた」とは記さない。日付差の理由や登録履歴の変更過程までは、今回確認した資料から確定できない。

診断名を当てる試験から、対応を考える試験へ

診断の正解が分かっても、何をいつ行うかは一つに決まらないことがある。本人の希望、経過観察へ戻れるか、費用、次の受診を確実に受けられるかで、適切な対応が変わる。Gohらはこのようなmanagement reasoning、すなわち治療・検査・その後の対応を組み立てる推論を評価した。[2]

先行研究との差は、単に難しい問題を増やしたことではない。正答の一つ選ぶだけでなく、複数の妥当な対応と、それを選ぶ条件を記述させた点にある。これはケア研究に近い問いだが、実際の患者との対話、本人の理解、生活への実行までは含まない。症例文に必要な情報が書かれている状況と、現場でその情報を聞き出す状況も違う。

採点表の中核と、点数の意味

各症例について、研究チームの一員、総合的に診る医師2人、関連する専門医2人の計5人が、修正Delphi法で採点表を作った。意見を繰り返し調整し、妥当とされた回答に点を与える仕組みである。二つの試行グループで確認し、採点表を修正してから用いた。[2]

3人の採点者のうち2人が、各回答を割付群を知らずに評価した。得点の差が10%を超えた場合は協議した。報告された採点の一致度は、統合した重

み付き κ で0.80だった。採点者が同じ答えへ似た点を付けられることは重要である。ただし、採点が一致することと、その得点が患者にとって良いケアを表すことは別である。

採点表は、妥当と認められる多様な内容を広く数える。40点だから不合格、60点なら安全という閾値はない。また、誤った提案に減点する方式ではなかった。正しい内容を多く挙げる回答が高得点になっても、同時に混ざる危険な内容を主要得点だけでは捉えきれない。著者ら自身も、この限界を明記している。

補足資料の例では、本人の希望やリスクの受け止め方、検査機器の費用や保険、退院後の情報の使い方も採点対象に入っている。したがって、医学知識だけを問う試験だったという批判は当たらない。一方、これらに言及することと、本人に確認して実行可能な計画を共同で作ることは同じではない。紙上の項目が含まれた段階から、ケアの成立へ飛躍しないことが必要である。

得点は上がった。しかし時間も増えた

評価	GPT-4支援群 / 対照群	原著の推定差と95%信頼区間
総合得点、100点換算	43.0 / 35.7	+6.5点 [2.7, 10.2]、 $p < 0.001$
対応方針の得点	40.5 / 33.4	+6.1点 [2.5, 9.7]、 $p = 0.001$
1症例回答の時間	801.5秒 / 690.2秒	+119.3秒 [17.4, 221.2]、 $p = 0.022$
総合得点、支援群対AI単独	43.0 / 43.7	-0.9点 [-9.0, 7.2]、 $p = 0.80$

表4。Table 2、要旨、本文に基づく。[2] 最後の行だけ比較相手がAI単独である。上3行は医師の二群の比較であり、同じ因果的な設計とみなさない。

43.0から35.7を引くと7.3であり、推定差6.5とは一致しない。これは直ちに誤植ではない。原著は参加者ごとの複数回答と症例ごとの難しさを考慮する混合効果モデルで差を推定している。単純平均の差とモデルの推定値は、別の量である。時間についても、単純平均の差111.3秒と、推定差119.3秒を区別する。

支援群はおおよそ2分多くかけて、得点が上がった。質を優先する課題としては価値があるが、同じ時間で多くの人へ対応できたという結果ではない。採点された完成回答は支援群の方が少なかった。未完了の回答は主要解析へ含まれないため、完成した回答の質と、限られた時間内の仕事全体の処理量を分けて読む必要がある。

この差をそのまま臨床の1患者当たり追加時間へ換算することもできない。実験の症例文、画面、質問数と、診療の流れは同一ではないからである。それでも、時間の増加を効率化という見出しで隠さないことには意味がある。導入後の評価では、入力する時間だけでなく、AI出力を読み、確認し、修正する時間を含めなければならない。

「差がなかった」から、何を言えるのか

AI単独の平均は43.7点で、支援群は43.0点だった。差の95%信頼区間は、支援群からAI単独を引く向きで-9.0から7.2点である。[2] この幅には、支援群が一定程度低い場合も高い場合も含まれる。差を検出できなかったことは、あらかじめ許容幅を決めて同等性を示したことではない。

さらに、AI単独は研究チームが検討した入力を用いた5症例各5回の出力である。医師が現場に近い使い方をした条件と、入力の作り方や反復単位が異なる。ここから、人を介さない診療の方が優れている、あるいは医師を省けるという結論は導けない。人が行う情報収集、説明、身体所見、実行、価値判断は、今回の点数にすべて含まれているわけではない。

ただし、比較の限界を述べて、AI単独の結果を無視するのも適切ではない。人が加わると何が変わるのかという問いを、研究課題として残す意味がある。重要な誤りを減らせたのか、本人の意向を反映できたのか、出力を適切に退けたのか。人の役割を守るためにも、何を付け加えたかを測れる形にする必要がある。

詳しい文章への加点と、重大な失敗を分ける

長さと時間を調整すれば解決するわけではない

Gohらは事後的な感度分析も行った。回答時間を調整した差は5.4点 [1.7, 9.0]、文字数を調整した差は3.7点 [0.7, 6.7]、両方を調整した差は3.3点 [0.4, 6.1] だった。[2] 主解析の結果が、単純に文章量だけで説明されるかを考える手掛かりになる。

しかし、時間や文字数はAIを使った結果として増えた可能性がある。介入後に変わった量を統計的に調整すれば、AIそのものの純粋な効果だけが残るとは限らない。時間をかけることが改善の経路なら、その一部を取り除くことになる。未測定 of 集中力などが時間と得点の両方に関われば、別の偏りも生じ得る。

したがって、追加解析を「長さも時間も関係なくAIが優れていた」と言い換えない。著者らはAIなしで考え直す時間を設ける比較なども今後の課題として挙げている。AIから新しい情報を得た効果と、いったん立ち止まって検討した効果を区別するには、設計段階から対照を用意する方が解釈しやすい。

害の評価には別の限界がある

この研究は、潜在的な害も探索的に調べている。群を知らされていない1人の評価者が、回答を読んで害の可能性と大きさを分類した。単位は症例全体ではなく、個々の回答である。実際に起きた患者の有害事象を数えたものではない。[2]

さらに、PMC著者原稿の本文と出版社のSupplement 6には数値差がある。例えば、支援群で「害の可能性が高い」とされた割合は本文3.8%、補足図4.2%であり、「重篤または死亡相当」の害の程度も本文7.6%、補足図7.7%と一致しない。補足図にはこの割合を再計算するための回答総数が示されておらず、今回の得点CSVだけでは照合できない。

本稿ではこの違いを未解決の報告上の不一致として残し、いずれかを確定した害発生率として用いない。出版社には訂正通知があるが、それは著者

の同等貢献・共同監督の表記に関するもので、これらの数値差を訂正した通知ではない。[4] 誤植と断定したり、著者に未確認の修正を代行したりしない。

数値差がなくても、一人の探索的な評定と、症例得点を主目的に設計された人数から、安全性の同等性を示すことはできない。平均点の改善と重大な誤りの減少は、別々に評価する必要がある。妥当な提案を一つ増やすことと、危険な提案を一つ残すことを同じ重みで相殺してよいかも、用途によって違う。

著者の結論と、Unknown Labの論評

Vaccaroらは、人とAIを組み合わせること自体を否定していない。単に組み合わせるだけでなく、どの仕事をどちらが担うかという過程と、時間、費用、重要な誤りを含む評価指標を研究する必要があると論じる。[1] Gohらは、複雑な症例課題でAIが医師の推論を支援する可能性を示しつつ、実際の診療で検証する必要があると結論する。[2]

Unknown Labがそこから取り出す論点は三つある。第一に、人を助けたかという問いには人単独との比較が必要だが、人と組む追加価値を語るにはそれだけでは足りない。第二に、文章の網羅性、危険な内容、人の判断、生活の結果を一系列の点数へまとめる前に分ける必要がある。第三に、独自の評価枠組みを導入するなら、一般的な整理や考え直しの支援より何が優れるのかを確かめる必要がある。

第三の論点は、二論文が直接比較した結果ではなく、本稿の研究設計上の提案である。例えば、詳しい指示を与えたAIと、短い指示だけのAIを比べれば、独自理論の効果と、単に必要な情報を明示した効果が混ざる。形式を整えた一般的な指示でも同じ改善が起こるなら、特別な枠組みの固有の効果을主張すべきではない。

観察・実験結果	言えること	まだ言えないこと
医師の無作為化された二群で得点差	この課題・利用条件で、GPT-4へのアクセスが完成回答の得点を改善したことを支持する。	実際の患者の安全、生活の改善、長期の学習効果。
支援群とAI単独で明確な差なし	限られた症例・反復で追加差を検出できなかった。	同等性、人を省けること、両条件の完全な公平性。
メタ分析の課題別・能力別の違い	効果が一様ではなく、次に比較する条件を示す。	課題や訓練を変えれば必ず同じ改善が生じるという因果効果。
害の探索的な評定	平均点だけでなく危険な回答を見る必要がある。	現実の有害事象率、安全性の同等性。

表5。原著の結果と、本稿で許容する推論の範囲。[1,2] 無作為化があっても、測っていない結果へ効果を延長することはできない。

二つの研究を読むうえで残る限界

三条件がある研究だけを選ぶことの代償

メタ分析は三条件を要求することで、追加価値を比べやすくした。しかし、人とAIが一緒でなければ実行できない仕事は除外され得る。人もAIも単独では完了できない複合的な課題まで、この平均値で否定することはできない。研究しやすい短い課題が多く選ばれるという偏りも残る。[1]

出版バイアスについても一括して「なし」としない。組合せによる追加的な優位では明確な非対称性を検出しなかった一方、人への支援効果では、良い結果の研究が公表されやすい可能性を示す検査結果があった。著者らは補正を行っていない。0.64という効果量を、偏りの可能性から切り離した確定値として扱わない。

症例の数と参加者の偏り

Gohらの375症例回答は、5種類の課題への繰り返し回答である。幅広い疾患、専門領域、言語、制度に対する外的妥当性を、その回答総数だけで支えることはできない。また、週1回以上LLMを使うと答えた医師は22人だった。熟練した利用者の上限性能でも、初めて使う人だけの効果でもない。[2]

最小84人という標本数の設計は、得点と回答時間を対象に、3人・13症例回答の予備データから行われた。重篤な誤りの差やAI単独との同等性を確かめるための設計ではない。主要評価で十分な人数でも、別の問いに十分とは限らない。

費用、人員、継続可能性

研究は監督者による支援と有償のChatGPT Plusを使う条件で実施された。参加謝礼は研修医100米ドル、attending 200米ドルだったが、これは実験への参加費用であり、医療機関の導入費用ではない。[2] 通常業務での教育、出力の確認、機器、契約、運用支援まで含めた費用対効果は測られていない。

時間が増えても、重要な誤りが十分に減るなら受け入れられる場合はある。反対に、得点が少し上がっても、確認する人員がなく実行できないなら、現場での価値は限られる。この判断には、誰の時間が増えるかを見る必要がある。医師の作業が減った代わりに本人や家族が説明・入力を引き受けた、という負担移転は今回の研究から分らない。

本稿は、これらの不足を理由に得点差を無価値とはしない。試験が答えた問いを認め、その先の導入判断には追加の証拠が要ると考える。研究段階ごとに測るべき結果を分ける方が、一つの試験へすべてを求めるよりも、次の検証を具体的にできる。

この研究からETIC 2.0が学べること

考え方と重なる点

技術を単体ではなく、人、仕事、環境との関係で評価するという考え方はETIC 2.0の問題意識と重なる。二論文は、人とAIが一緒にいるという配置だけでは成果を予測できず、課題、利用過程、比較基準が重要であることを示す。ただし、原著はETIC 2.0を評価しておらず、この一致を理論の実証と呼ぶことはできない。

とくに重要なのは、本人にとっての意味ある生活と、専門家が採点しやすい出力を同一視しないことである。適切な選択肢を多く挙げた文章でも、本人が実行できない、選びたくない、重要な条件が抜けている場合がある。一方、短い回答でも、その人が次に何を確かめるかを明確にするなら価値があり得る。どちらが優れるかは、例の説得力ではなく比較で確かめる必要がある。

答えるべき問い

ETIC 2.0に基づいて説明や検討の順序を整えると、どの見落としが減るのか。その改善は、一般的なチェック項目を与えるだけでも生じないか。回答が詳しくなる代わりに、読む負担や重要な誤りが増えていないか。これらに答えなければ、枠組みを用いた文章の完成度と、枠組み自体の追加価値を取り違える。

また、AIの出力を比較する研究と、その出力を人が読んで判断する研究は分けなければならない。前者で重要な条件が多く含まれたとしても、後者で人が条件を理解し、誤りを退け、適切に相談するとは限らない。読みやすい文章ほど誤った内容も受け入れられる可能性は、別の人を対象とする検証課題である。

具体的な見直し案

次の表はUnknown Labの未検証の提案である。新しい評価項目を際限なく足すのではなく、何を主張したいかに対応する比較と指標を選ぶ。既存の簡単な整理で足りるなら、独自の手順を減らす方向も含める。

確かめる問い	比較に必要な条件	主に見る結果
AIの回答を整える方法に追加価値があるか	同じモデル・入力条件で、通常指示、一般的な構造化指示、検討する方法を比較	重要な見落とし、危険な提案、妥当な内容。長さも記録する。
AIは人の判断を助けるか	人が通常の資料を使う条件と、AIを使える条件を比較	最終判断、誤りの修正、時間、確信と正確さのずれ。
人と組む意味はどこにあるか	同じ課題と評価でAI単独も参照し、入力条件の差を記録	平均点だけでなく、人が除いた重大な誤りと加えた判断。
現実のケアへ移せるか	対象を限定し、実施体制と倫理的手続を整えた別研究	本人の結果、実行可能性、主体別の負担、継続性。

表6。公開研究から導いた比較設計の候補。新しい実験を実施した結果でも、一般的な臨床導入基準でもない。人を対象とする研究は、出力だけの比較とは別の手続と確認を要する。

公平さは、すべてを機械的に同じ長さへ切ることではない。実用条件での総合的な違いを見たいのか、説明の量を揃えて内容の違いを見たいのかで、適切な設計は変わる。前者なら時間や費用を成果と併せて測る。後者なら出力上限や指示の詳しさを事前に揃える。結果を見てから有利な方へ切り替えないことが重要である。

さらに、主評価には、事前に定義した重大な失敗の有無を置くことを検討する。何を重大とするかは専門職と当事者の視点を含めて決め、平均得点で埋め合わせない。採点者には条件名を伏せ、不一致がどこで起きたかを残す。これはAIに有利な結果を作るためではなく、見栄えの良い出力を過大評価しないための方法である。

どんな結果なら、見直し案を取り下げるのか

提案は、期待した結果が出ない場合の判断まで含める必要がある。「効果が見えなければ、もっと詳しい枠組みにする」とだけ考えると、方法そのものを検証できなくなる。

得られた結果	必要な判断
一般的な構造化指示と差がなく、負担も同程度	独自の枠組みに固有の効果という主張を保留する。簡単な方法を優先する。
平均点は上がるが重大な失敗は減らない	安全性の改善とは呼ばない。得点と失敗を別に報告する。
重要な内容は増えるが、読む時間と誤採用も増える	網羅性だけを成果にしない。情報量や提示方法を見直す。
出力は改善するが、人の最終判断は改善しない	出力評価の範囲へ主張を戻す。利用過程を別途調べる。
人と組んでも平均点は上がらないが重大な誤りを減らす	平均点だけで人の役割を否定しない。便益と必要資源を別に評価する。

表7。Unknown Labの反証・縮小条件。今回の二論文からこれらの結果がすでに得られたという意味ではない。

差がないという結論自体にも、十分な精度が要る。小さな試験の非有意結果だけで方法を同等と認定しない。意味のある最小の差、許容できない失敗、必要な人数・反復を事前に検討し、判断できない場合は判断できないと報告する。この留保は、望む結果が出なかった方法だけでなく、自分たちの方法にも同じように適用する。

次に読む論文

次は、Sidenらの「A typology of physician input approaches to using AI chatbots for clinical decision-making」を読む。[5] AIに何を入力したかを、医師22人への面接と、二つの無作為化試験の利用ログから調べた混合研究法の研究である。面接は2024年5月から7月に行われ、Framework Methodを用いた分析から、症例全体の貼り付け、一部の貼り付け、要約、検索に近い質問という入力の分類を作っている。

推薦理由は、AIが使えるという割付と、実際にどう使ったかの間を調べる点にある。得点差だけでは、どの情報を渡し、何を省き、出力をどう判断したかが見えない。この過程を調べることは、次の比較条件を具体化する材料になる。

ただし、利用ログには今回のGohらの試験のデータも含まれる。同じデータを使った後続分析を、独立した再現試験として証拠の数へ足してはならない。また、入力の型は無作為に割り付けられていない。入力方法と成績に明確な関連がなかったことを、「どんな聞き方でも同じ」と読むこともできない。面接で分かった使い方と、効果の比較を分けて読む予定である。

結論

AIの支援で人の回答の得点が上がることには、意味がある。しかし、その結果だけで人とAIの組合せが最良だとも、重大な誤りが減ったとも、実際のケアが良くなったとも言えない。今回の二論文は、比較相手と評価指標を明確にすると、何が改善し、何が未確認かを分けられることを示している。

Unknown Labが次に進めるべきなのは、もっと多くの項目を含む回答を作ることだけではない。一般的な方法でも得られる改善を除き、重要な見落としや危険な提案が減るかを確かめることである。そこから人の判断、さらに生活の結果へ進むときには、別の証拠が必要になる。良い文章を作れたことを認めながら、良い判断を支えたかは問い続ける。その区別を、研究の設計に戻したい。

参考文献

1. Vaccaro M, Almaatouq A, Malone T. When combinations of humans and AI are useful: A systematic review and meta-analysis. Nature Human Behaviour. 2024;8:2293-2303. [doi:10.1038/s41562-024-02024-1](https://doi.org/10.1038/s41562-024-02024-1). [公開全文](#)。本文、Methods、補足資料を確認。
2. Goh E, Gallo RJ, Strong E, et al. GPT-4 assistance for improvement of physician performance on patient care tasks: a randomized controlled trial. Nature Medicine. 2025;31:1233-1238. [doi:10.1038/s41591-024-03456-y](https://doi.org/10.1038/s41591-024-03456-y). [PMC著者原稿全文](#)、出版社の表・図説明、補足資料、公開Source Dataを確認。出版社の本文全体を購読版で確認したものではない。
3. Stanford University. Management Reasoning With AI Chat Bots. ClinicalTrials.gov, NCT06208423. [試験登録](#)。2026年9月25日確認。論文と登録情報の実施日を区別。
4. Goh E, et al. Publisher Correction: GPT-4 assistance for improvement of physician performance on patient care tasks: a randomized controlled trial. Nature Medicine. 2025. [doi:10.1038/s41591-025-03586-x](https://doi.org/10.1038/s41591-025-03586-x)。著者の同等貢献・共同監督表記に関する訂正。
5. Siden R, Kerman H, Gallo RJ, et al. A typology of physician input approaches to using AI chatbots for clinical decision-making. npj Digital Medicine. 2026;9:14. [doi:10.1038/s41746-025-02184-y](https://doi.org/10.1038/s41746-025-02184-y)。オンライン公開2025年12月5日。全文を確認し、次に読む論文として推薦。