「人が最後に確認する」だけで安全か

AI診断支援における信頼較正と人間監督の限界

Sakamotoら（2024）を読む

森元大樹 / Unknown Lab / ORCID 0009-0009-6444-750X
論評記事 Version 1.0 2026-07-18
DOI: 10.5281/zenodo.21429496

要旨

医療やケアでAIを利用するとき、『最後は人が確認する』という安全策がしばしば置かれる。Sakamotoらは日本の医師20人を対象に、AIの上位10鑑別に正解が含まれるかを症例ごとに確認させる介入を検討した。診断正確性は介入群41.5%、対照群46.0%で、有意な改善は認められなかった。一方、AIリストの正誤を正しく判断できた症例では診断も正しい傾向があったが、これは介入の因果効果を示すものではない。

本稿は、広い確認質問を加えることと、確認を実行できる条件を設計することを区別する。ETIC 2.0への含意として、人間責任を最終承認の役割だけでなく、独立した見立て、具体的な矛盾確認、保留と相談の権限、事後監査を実行できるケア生態系として具体化する必要性を論じる。

主要結果

41.5%	46.0%	p=.27	61.5%
確認を求めた介入群	確認を求めない対照群	群間に有意差なし	AIリストの正誤判断

形式的な確認から、確認可能な協働へ

形式的な人間監督	確認可能な協働
AIを見た後で『正しいか』と聞く	AIを見る前の見立ても残す
AIの候補だけを示す	根拠・欠けた情報・適用外条件を並べる
Yes / Noで承認する	採用・修正・却下・保留・相談を選べる
人の注意力に任せる	矛盾点を具体的に問い、確認負担を下げる
最終判断だけを記録する	依存の仕方と生活上の結果を監査する

English Summary

Sakamoto et al. studied 20 physicians who reviewed 20 clinical cases with an AI-generated top-10 differential diagnosis list. Prompting physicians to judge whether the correct diagnosis was present in the AI list did not improve diagnostic accuracy (41.5% vs 46.0%; p=.27). Correct judgments about the AI list were associated with higher diagnostic accuracy, but this within-group association does not establish a causal effect of the intervention.

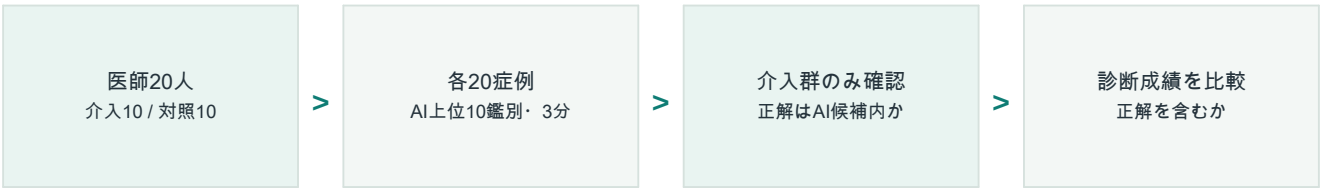
This critical commentary distinguishes formal human review from conditions that make verification possible: an independent initial assessment, information about AI performance and limitations, specific contradiction checks, authority to defer or escalate, and post-use audit. It is an unreviewed commentary and not evidence that ETIC 2.0 is effective.

Keywords: trust calibration, human oversight, artificial intelligence, diagnostic decision support, human-AI collaboration, occupational therapy, ETIC 2.0

30秒で分かる研究の要点

項目	内容
研究	日本の総合診療系医師20人による2群比較。各医師が20症例を判断
介入	「正解がAIの上位10鑑別に含まれるか」を症例ごとに確認
主要結果	介入群41.5%、対照群46.0%。有意差なし（p=.27）
もう一つの結果	確認判断の正確性は61.5%。正しく確認できた症例ほど診断も正確だった
読み方	広い確認質問だけでは成績は上がらない。ただし、確認を支える設計まで無効と示した研究ではない

研究の流れ



研究が試したのは、最小限の「確認」だった

対象となったのは、獨協医科大学病院の診断・総合診療部門に所属していた医師20人です。研究は2023年8月9日から9月25日まで行われ、医師は介入群10人と対照群10人に割り付けられました。論文は準実験と表記していますが、割付には卒後年数で層別したコンピューター無作為化が使われ、参加者には自分の群が知らされませんでした[1]。

20の症例は、長野中央病院を受診した実患者のAI問診記録に基づきます。元データは2019年5月から2022年4月までの381症例で、そこから疾患領域、病気の頻度、症状の典型性を考慮して20症例が選ばれました。AIの上位10鑑別に正解が含まれる症例と含まれない症例は10件ずつです。

両群の医師は、AIがまとめた病歴と上位10件の鑑別診断を読み、1症例3分以内に1〜3個の診断候補を書きました。違いは一つだけです。介入群には、正解がAIの候補に含まれているかを検討するよう促し、YesかNoで回答させました。対照群には、この確認を求めませんでした。

研究のキモは、新しい診断AIを作ったことではありません。AIの答えを評価するメタ判断を、診断手順に一つ加えたことです。

先行研究から、何を一歩進めたのか

同じAI問診を使ったHaradaらの2021年の研究では、医師22人が16症例を判断しました。AIの鑑別リストを見た群と見なかった群で、診断正確性に差はありませんでした（57.4%対56.3%、絶対差1.1ポイント、95%信頼区間 -9.8〜12.1、p=.91）。ただし、医師の回答はAI候補に引き寄せられていました[2]。

Jabbourらが457人の臨床家を対象にした研究では、正しいAIは診断を助ける一方、偏ったAIは正確性を低下させました。しかも、AIに説明を付けても害は十分に減りませんでした[3]。

これらの研究は、AIの出力が人の判断を動かすことを示しています。Sakamotoらが一歩進めたのは、医師が症例ごとにAIリストの正誤を見分けられたかを測り、その判断と最終診断を結び付けたことです。

ただし、非医療分野の先行研究で用いられた「信頼較正」とは違いがあります。OkamuraとYamadaのドローン実験では、システムが利用者の過信を検知し、その時点で警告を出しました[4]。Sakamotoらの介入には、AIの性能情報も、症例ごとの不確実性も、過信への警告もありません。検証されたのは、支援機能を備えた信頼較正システムではなく、広い確認質問を一つ追加する方法です。

医療を含む自動化バイアス研究の体系的レビューも、元の情報と自動化された助言を照合する作業の複雑さが、誤った助言の見逃しに関係しうると整理しています[5]。確認を求めることと、確認しやすい条件を整えることは同じではありません。

結果は『確認すれば良い』とは言わなかった

主要評価項目は、医師の診断候補に最終診断が含まれた割合です。

- 介入群: 平均8.3/20問、41.5%
- 対照群: 平均9.2/20問、46.0%
- 生の差: 介入群が4.5ポイント低い
- 論文が示した群間差の95%信頼区間: -0.75〜2.55
- p=.27

統計学的な有意差はありませんでした[1]。確認を加えても改善しなかったという結果です。ただし、介入が害を与えたと結論することも、確認という行為が常に無効だと結論することもできません。この研究は15ポイントの改善を検出する想定で設計されており、より小さな効果を否定できる規模ではありません。

次に、介入群の医師が「AIリストに正解があるか」を正しく判断できた割合を見ます。

- 確認判断が正しかった: 123/200件、61.5%
- 確認判断が正しい場合の診断正確性: 67/123件、54.5%
- 確認判断が誤った場合の診断正確性: 16/77件、20.8%
- 調整オッズ比: 5.90、95%信頼区間2.93〜12.46、p<.001

確認判断の正しさと診断正確性には強い関連がありました。しかし、これは「確認を促したから診断が良くなった」という因果効果ではありません。正しい診断を知っている医師ほど、AIリストにその診断があるかも正しく判断しやすいからです。確認能力が診断能力の一部を測っている可能性があります。

さらに、介入群が報告したAIへの信頼度は72.5%、対照群は45.0%でした（p=.12）。選ばれた20症例でAIリストに正解が含まれた割合は50%です。確認を求めたことが、かえってAIへの注意や期待を高めた可能性もありますが、群間差は有意ではなく、ここから断定はできません。

この研究の強みと、慎重に読むべき点

強みは、AI単体の正確性ではなく、AIを見た人間の最終判断を主要評価項目にしたことです。実患者の病歴から症例を作り、AIが正しい場合と誤る場合を半数ずつ含め、医師がAIの正誤を見分ける課題を明示しました。人間-AIチームを評価しようとした研究です。

一方、限界は少なくありません。

- 医師20人、単一部門、若い総合診療系医師が中心です。
- 実患者をその場で診療した研究ではなく、1症例3分の文章課題です。
- 介入群はAIの事前性能や、AIが外しやすい条件を知らされていません。
- 正解がAIリストにあるかを判断する作業は、診断そのものと同じくらい難しい可能性があります。
- 介入群内の回帰分析について、同じ医師が20症例を判断した反復測定の調整方法は明記されていません。オッズ比の精度は慎重に読む必要があります。
- 患者アウトカム、医師の満足、実際の業務負担は評価されていません。

著者ら自身も、この方法が十分な信頼較正ではなかった可能性を認めています。次の研究では約158人を想定し、AIの正確性を事前に知らせ、過信や過小信頼を検知したときに警告を出す仕組みを加えるとしています。また、「AIは正しいか」ではなく、「どの所見と矛盾するか」を考えさせる具体的な振り返りを候補に挙げています[1]。

ETIC 2.0が受け取るべきメッセージ

ここから先は、Sakamotoらが直接検証した結論ではなく、Unknown Labによる設計上の考察です。

ETIC 2.0は、人間の責任を重視します。しかし、「人が最後に確認した」という記録だけでは責任を果たしたことになりません。確認する人に、判断材料、時間、権限、相談経路がなければ、最終確認は儀式になります。

形式的な確認を、確認可能な協働へ変えるには、少なくとも次の違いが必要です。

形式的な人間監督	確認可能な協働
AIを見た後で「正しいか」と聞く	AIを見る前の見立ても残す
AIの候補だけを示す	根拠、欠けた情報、適用外条件を並べる
Yes/Noで承認する	採用、修正、却下、保留、相談を選べる
人の注意力に任せる	矛盾点を具体的に問い、確認負担を下げる
最終判断だけを記録する	AIへの依存と、その後の生活上の結果を監査する

作業療法で考えると、AIが転倒リスクや運動プログラムを示したときに、「この提案は正しいですか」と専門職へ聞くだけでは足りません。本人が大切にしている外出、疲労の変化、住環境、家族の負担、利用できる地域資源と矛盾していないかを、具体的に照合できる必要があります。

この研究は、ETIC 2.0の有効性を証明するものではありません。しかし、「人間責任の保持」を役職名や最終クリックではなく、責任を実行できるケア生態系の設計として具体化すべきだと教えています。

次に読む論文

次に読みたいのは、Gohらの無作為化臨床試験です[6]。

米国の医師50人を、GPT-4を使える群と通常の資料だけを使う群に分け、最大6症例の診断推論を評価しました。診断推論得点は76%対74%、調整差2ポイント（95%信頼区間 -4〜8、p=.60）で、LLMを使えることによる有意な改善はありませんでした。一方、LLM単独は通常資料群より16ポイント高い得点でした（95%信頼区間2〜30、p=.03）。

能力の高いAIがそこにあっても、人間-AIチームが自動的に強くなるわけではない。この論点は、Sakamotoらの研究から次へ進むための自然な問いです。

結論

Sakamotoらの研究では、AIの鑑別リストに正解があるかを確認するよう促しても、医師の診断正確性は改善しませんでした。確認判断自体も61.5%にとどまりました。

必要なのは、最後に人を置くことだけではありません。AIを見る前の独立した見立て、AIの性能と限界、具体的な矛盾確認、保留と相談の権限、結果の監査までを一つの仕組みとして設計することです。

人間を安全装置と呼ぶなら、その人が安全装置として働ける条件まで設計しなければなりません。

参考文献

1 Sakamoto T, Harada Y, Shimizu T. Facilitating trust calibration in artificial intelligence-driven diagnostic decision support systems for determining physicians' diagnostic accuracy: quasi-experimental study. JMIR Form Res. 2024;8:e58666. doi:10.2196/58666 (https://doi.org/10.2196/58666)

2 Harada Y, Katsukura S, Kawamura R, Shimizu T. Efficacy of artificial-intelligence-driven differential-diagnosis list on the diagnostic accuracy of physicians: an open-label randomized controlled study. Int J Environ Res Public Health. 2021;18(4):2086. doi:10.3390/ijerph18042086 (https://doi.org/10.3390/ijerph18042086)

3 Jabbour S, Fouhey D, Shepard S, et al. Measuring the impact of AI in the diagnosis of hospitalized patients: a randomized clinical vignette survey study. JAMA. 2023;330(23):2275-2284. doi:10.1001/jama.2023.22295 (https://doi.org/10.1001/jama.2023.22295)

4 Okamura K, Yamada S. Adaptive trust calibration for human-AI collaboration. PLoS One. 2020;15(2):e0229132. doi:10.1371/journal.pone.0229132 (https://doi.org/10.1371/journal.pone.0229132)

5 Lyell D, Coiera E. Automation bias and verification complexity: a systematic review. J Am Med Inform Assoc. 2017;24(2):423-431. doi:10.1093/jamia/ocw105 (https://doi.org/10.1093/jamia/ocw105)

6 Goh E, Gallo R, Hom J, et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. JAMA Netw Open. 2024;7(10):e2440969. doi:10.1001/jamanetworkopen.2024.40969 (https://doi.org/10.1001/jamanetworkopen.2024.40969)

本稿は公開済み査読論文を批判的に読み解いた未査読の論評記事であり、ETIC 2.0の有効性を検証した研究ではない。個別の診断、治療、リハビリテーションを指示するものではない。

執筆: Unknown Lab 編集部（AI支援） / 確認: 作業療法士