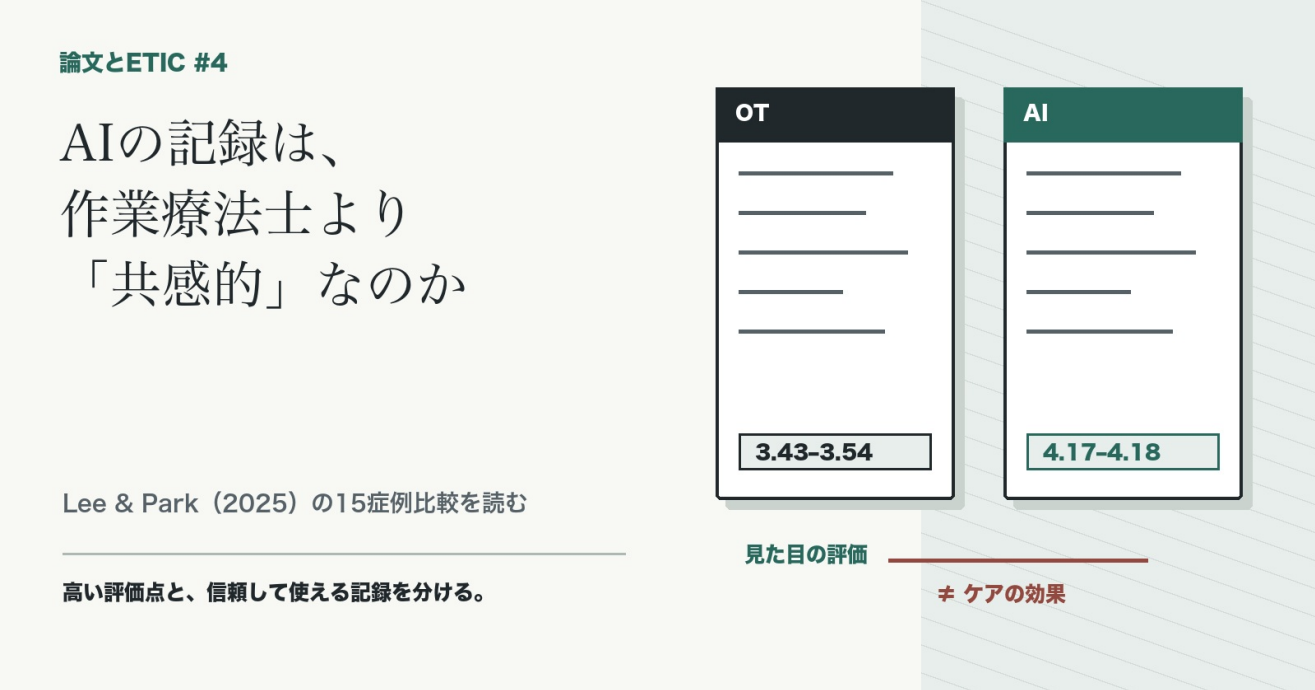
AIの記録は、作業療法士より「共感的」なのか

Leeら（2025）の15症例比較研究を読む



Daiki Morimoto / Unknown Lab / ORCID 0009-0009-6444-750X
論評記事 Version 1.0 2026年8月13日
DOI: 10.5281/zenodo.21897214

要旨

Leeらは15の標準化された作業療法症例を用い、作業療法士とChatGPT-3.5が同じS/O情報から作成したA/P記録を比較した。記録は作業療法士5人と患者・介護者5人が盲検評価し、AI文書は質3項目と共感3項目のすべてで高い平均点を得た（すべてp<0.001）。

ただし、測定されたのは架空症例の文書がどう見えたかであり、事実単位の正確性、記録時間、患者成果、費用、安全性は検証されていない。本文と表には相関項目の不一致と成立しないICC信頼区間もある。本稿は、読みやすく共感的に見える文書と、本人に応答する安全なケアを分けて評価する。

論文とETIC #4 Version 1.0公開版 / 2026年8月13日

記録を書く時間を減らし、その分を本人との対話に使えるなら、AIは作業療法の助けになるかもしれません。

一方、整った文章が短時間で出てくると、私たちは内容まで正しいように感じやすくなります。本人の生活目標が抜けた記録でも、専門用語と配慮ある表現がそろっていれば、「質が高い」「共感的だ」と評価してしまう可能性があります。

Leeらが2025年に発表した研究は、15の標準化された作業療法症例について、作業療法士とChatGPT-3.5が作成した記録を比較しました。作業療法士と患者・介護者が評価すると、AI文書は質と共感の全6項目で人間文書より高い平均点でした[1]。

ただし、この研究が測ったのは、架空症例から作った文書がどう見えたかです。AIが患者を理解したか、事実誤認を減らしたか、専門職の時間を節約したか、本人の作業参加を改善したかは測っていません。

さらに原文を表まで確認すると、本文と表の相関ペアが一致せず、信頼区間にも成立しない数値があります。高い評価点と、信頼して使える記録は同じではありません。

30秒で分かる研究の要点

項目	内容
研究	標準症例を使った盲検比較横断研究
症例	架空の作業療法症例15件
人の参加者	記録作成OT5人、評価OT5人、患者・介護者5人
比較	同じ症例情報から作った人間のA/P記録とChatGPT-3.5のA/P記録
評価	質3項目、共感3項目を5段階評価
主結果	AI文書の平均評価が全6項目で高い。すべてp<0.001
読み方	文書の知覚評価。正確性、時間短縮、患者成果を実証した研究ではない

English Summary

Lee and Park compared Assessment and Plan notes written by five occupational therapists with notes generated by ChatGPT-3.5 for 15 standardized fictional cases. Five occupational therapists and five patients or caregivers rated the blinded documents. AI notes received higher mean ratings across three quality and three empathy domains (all p<0.001). The study measured perceived document quality, not fact-level accuracy, time savings, safety, patient outcomes, or causal effects of implementation. Internal inconsistencies in the reported correlations and an impossible confidence interval also require caution.

Keywords: occupational therapy, clinical documentation, artificial intelligence, ChatGPT, empathy, human oversight, ETIC 2.0

本稿の位置づけ

本稿は公開済み査読論文をUnknown Labが批判的に読み解いた未査読の論評記事である。著者の結論、研究から直接分かる観察結果、因果効果の可否、Unknown Labの論評、ETIC 2.0への支持・挑戦・更新案を分けて記載している。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。

補強記事05から、何を一歩深めるのか

直前の補強記事「作業療法でAIはどこまで使われているのか」は、作業療法固有のAI研究がまだ少なく、リハビリテーション全体でも外部検証や説明可能性が不足していることを整理しました。技術が動くことと、日常臨床で安全に役立つことの間には距離があります。

その記事で紹介したDeğerliらは、111人の作業療法士とChatGPT-3.5に、3症例についてモデル、評価、介入の選択肢を答えさせました。1症例では一部の選択が似ていましたが、他の2症例では違いがありました[3]。これは、AIが作業療法の言葉を使えるか、選択が専門職集団と似るかを調べた研究です。

Leeらは問いを記録文書へ進めました。A/Pとは、SOAP記録のうちAssessment、Planに当たる部分です。得られた情報をどう解釈し、何を次の支援として計画するかを書くため、単なる要約ではありません。

先行するAyersらは、公開医療相談サイトの患者質問に対する医師とChatGPTの回答を比較し、AI回答が質と共感で高く評価される場合があると報告しました[2]。Leeらは一般医療相談ではなく、作業療法の記録へ対象を絞りました。さらに質と共感を各3項目へ分け、専門職だけでなく患者・介護者も評価者に含めました。

新しさは、「AIがそれらしい答えを出したか」ではなく、作業療法の文書として、どのように受け取られたかを複数の視点で評価したことです。

誰が、何を評価したのか

研究は韓国で行われた、盲検化された比較横断研究です。実際の患者記録ではなく、韓国で出版された作業療法症例集の架空症例を使いました[1]。

人の参加者は15人でした。

役割	人数	研究で行ったこと
記録を作成した作業療法士	5人	症例のS/OからA/Pを作成
記録を評価した作業療法士	5人	人間文書とAI文書の質・共感を評価
患者または介護者	5人	同じ文書の質・共感を評価

作業療法士は免許を持ち、担当領域の臨床経験がありました。参加者は専門職ネットワークや患者支援チャネルから目的抽出されました。人口統計情報は収集されていません。

ここでいう患者・介護者は、15症例の本人や家族ではありません。架空症例の文書を読み、受け取られ方を評価した生活経験者です。患者側の視点を入れた点は重要ですが、実際の支援関係で「自分のことを理解してもらえた」と評価したわけではありません。

症例数の記載には解けない曖昧さがある

原文は15症例を無作為に選んだと書いています。その直後に、筋骨格11、発達10、神経10、心理社会6と記し、合計すると37になります[1]。

これは症例集全体の候補数なのか、選択後の症例に複数領域が重なっているのか、本文だけでは判断できません。15症例の実際の領域構成は確定できないため、本稿では「15症例」とだけ記し、領域別の一般化に使いません。

人とAIの文書を、どうそろえたのか

各症例にはSubjectiveとObjective、つまり本人から得た主観情報と観察・測定した客観情報に相当する材料がありました。人とAIは同じ情報からA/Pを作りました。

ChatGPT-3.5への入力は2025年4月28日と29日に行われました。症例ごとに別セッションを使い、前のやり取りの影響を避けました。作業療法士とAIには同じ標準記録ガイドラインを示し、文書はA4約半ページ、12ポイントにそろえました[1]。

その後、研究者は生成者が分からないよう匿名化し、文体差を除き、用語を統一しました。評価者には文書を無作為な順序で提示しました。

この標準化には二つの面があります。

一つは、AI文書が長いだけで高く評価されることを防ぎ、同じ条件で比べやすくしたことです。もう一つは、研究者が文体と用語を整えたことで、実際に生成された文書そのものから変わった可能性があることです。どの部分を、誰が、どの基準で修正したかは十分に報告されていません。

したがって、これは「現場で人とAIがそのまま書いた文書」の比較ではなく、比較のために研究者が整えた文書の比較です。

質と共感を、各3項目で測った

評価者は各文書を5段階で採点しました。

大項目	下位項目	研究上の意味
質	完全性	必要な情報が含まれているか
質	正確性	事実・臨床的に正しいと評価されるか
質	整合性	作業療法の実践基準と合うか
共感	認知的共感	本人の状態を理解しているように見えるか
共感	感情的共感	配慮ある肯定的な態度が表れるか
共感	行動的共感	明確で応答的な伝え方になっているか

共感の考え方はCARE MeasureとJefferson Scaleを理論的な基礎にしました[4][5]。ただし、これらは対面での相談や臨床家に対する患者の知覚を評価するために開発された尺度です。Leeらは尺度をそのまま使わず、書面評価のための3項目へ翻案しました。

著者自身も、書面の質と共感を測る標準尺度が存在しないことを限界に挙げています。5段階の「正確性」が高いことは、原情報と一文ずつ照合して誤りが少なかったことを意味しません。評価者が正しいと感じたことを示します。

どのように差を検証したのか

人間文書とAI文書の平均差には、対応のあるt検定を使いました。質と共感が一緒に高くなるかにはPearson相関、評価者がどの程度一致したかには二元配置変量効果・絶対一致の級内相関係数、ICCを使いました。統計的有意水準は両側p<0.05でした[1]。

ここには重要な分析上の問題があります。

同じ15症例を同じ評価者が繰り返し採点したため、評価は互いに独立ではありません。症例の難しさ、評価者の厳しさ、記録作成者の書き方が結果へ影響します。このようなデータは、評価が症例や人に入れ子になった多層構造です。

しかし原文は、単純な対応のあるt検定と相関を使い、症例・評価者・作成者のクラスタリングを考慮した混合効果モデルを報告していません。評価数を独立した観察のように扱っていれば、標準誤差が小さくなり、p値が必要以上に小さくなった可能性があります。これはUnknown Labによる方法上の批判であり、原論文が検討した感度分析ではありません。

AI文書の平均評価は高かった

結果として、質3項目と共感3項目のすべてで、AI文書の平均評価が人間文書より高く、すべてp<0.001でした[1]。

結果	人間文書	AI文書	注意
質・OT評価の平均範囲	3.43〜3.54	4.17〜4.18	各項目の完全な平均・SDは本文表にない
共感の平均	約3.6	4.0超	OTと患者・介護者の双方で同方向
良い/非常に良い評価	基準	2.7〜2.8倍	記述比。95% CIなし
共感的/非常に共感的評価	基準	2.6〜3.0倍	記述比。95% CIなし
3未満の低評価	同程度	同程度	AIで低評価が消えたわけではない

AI文書は平均では「良い」に相当する4点を超え、人間文書は「受け入れ可能」に相当する3点台でした。高評価の割合もAIで大きく見えました。

一方、原文は平均差の効果量や95%信頼区間を報告していません。統計的に差があることと、実務で重要な差であることを分けて判断できません。評価尺度自体も未検証なので、0.6点ほどの差が安全や生活に何を意味するかは不明です。

高い評価点でも、質と共感は連動しなかった

研究は、質の高い文書ほど共感も高く評価されるかを調べました。

人間文書では、質3項目と共感3項目の9組すべてに正の相関があり、r=0.290〜0.457、すべてp<0.001でした。完全性、正確性、整合性が高く評価された文書は、認知的・感情的・行動的共感も高く評価される傾向がありました[1]。

AI文書では、9組の多くが-0.084〜0.064で、有意な相関はほぼありませんでした。AIは質と共感の平均点をそれぞれ高く出しましたが、同じ文書の中で二つが一緒に高くなるとは限りませんでした。

この結果は、AIが「共感の言葉」を質の高い臨床推論と統合できていない可能性を示唆します。ただし、相関がない理由は、AIの点数が全体に高く範囲が狭い天井効果でも起こり得ます。相関だけから心理的な共感能力を推定することはできません。

本文と表1は、有意だった組を違って記している

原文本文は、AI文書で唯一有意だった組を「正確性と感情的共感、r=0.148、p=0.005」と記しています。しかし表1でr=0.148なのは「正確性と認知的共感」です。感情的共感と正確性はr=-0.014で、有意印はありません[1]。

本稿は表1の配置を優先しますが、著者または出版社による訂正は確認できていません。したがって、「AIでは正確性と認知的共感に弱い相関が一つ報告されたが、本文の項目名と不一致」と記します。

評価者間信頼性にも、読み飛ばせない誤植がある

ICCは、評価者の採点がどの程度一致したかを0〜1で表します。値が高いほど一致が強いと読みますが、用途や評価者数により解釈は変わります。

領域	人間文書 ICC（95% CI）	AI文書 ICC（95% CI）
完全性	0.742（0.655〜0.808）	0.672（0.560〜0.756）
正確性	0.598（0.462〜0.700）	0.527（0.365〜0.647）
整合性	0.580（0.438〜0.687）	0.743（0.654〜0.809）
認知的共感	0.718（0.622〜0.790）	0.694（0.590〜0.772）
感情的共感	0.754（0.670〜0.817）	0.576（0.431〜0.684）
行動的共感	0.615（0.485〜0.713）	0.551（原文CIは使用不能）

著者は、人間文書の方が全体として評価者間の一致が高かったと解釈しました。6項目中5項目では人間文書が同等以上ですが、整合性ではAI文書のICC 0.743が人間の0.580を上回ります。「人間が全項目で高い」わけではありません。

もっと重大なのは、AI文書の行動的共感です。原文はICC 0.551、95% CI 0.698〜0.665と記しています。下限が上限より大きく、点推定0.551も区間に入りません。統計的に成立しない表記なので、本稿はこの信頼区間を使用しません。

表の誤植は研究全体を直ちに無効にしません。しかし、AI文書の正確性を論じる記事で、原論文自身の表に未解決の不整合があることは重要です。論文もAI出力も、「査読済み」「整っている」という外観だけでなく、数値を追跡して確認する必要があります。

著者の結論、観察結果、因果効果を分ける

著者の結論

著者は、標準症例ではAI文書が質と共感で高く評価され、作業療法士が確認・編集する下書きとして記録負担を減らす可能性があるとしました。同時に、AI文書は評価者による解釈のばらつきがあり、人間監督、文脈への適応、実臨床での検証が必要だと述べました[1]。

研究から直接観察されたこと

比較条件をそろえた15症例の文書を、作業療法士5人と患者・介護者5人が評価したとき、AI文書の平均点が高かった。人間文書では質と共感の点数が一緒に動き、AI文書ではほぼ連動しなかった。評価者間一致は項目により異なった。

因果効果として言えないこと

研究は、AI導入あり・なしを現場へ無作為に割り付けた介入試験ではありません。次の効果は分かりません。

- 記録時間や残業が減るか。
- 誤記、幻覚、監査指摘、事故が減るか。
- 専門職が本人と話す時間が増えるか。
- 本人が理解され、選択できたと感じるか。
- 支援計画が本人の作業目標に近づくか。
- 健康、生活、社会参加が改善するか。

AI文書が高く評価されたことを、「AIを使えばケアが改善する」という因果効果へ拡張してはいけません。

Unknown Labの批判的論評

1.「正確性」は、事実監査ではない

評価者は文書を5段階で採点しました。しかし、症例の原情報と文書の各記述を照合し、追加された事実、抜けた情報、矛盾、危険な提案を数えたわけではありません。

専門用語が自然で、計画がもっともらしければ、誤った推論も高く評価される可能性があります。実装前に必要なのは「正しそうか」ではなく、事実単位の一致率、重大誤り、修正が必要な文、見逃しやすい誤りを測ることです。

2.「共感的に見える」と「共感的なケア」は違う

共感は文章中の優しい言葉だけではありません。本人の視点を理解し、言葉になりにくい経験を確かめ、選択を尊重し、応答を変える関係の過程です。

この研究の患者・介護者は架空症例の文書を読みました。文書の受け取られ方を評価した点は有意義ですが、自分の語りが正しく残ったか、自分の目標に計画が応じたか、実際のやり取りで尊重されたかは測れません。

AIは配慮ある表現を安定して生成できます。それは文書の可読性を高める可能性があります。しかし、共感らしい文を生成できることと、本人との関係の中で応答責任を負うことは別です。

3. 人間側の比較条件が十分に見えない

記録作成OT5人が15症例をどう分担したか、臨床経験年数、普段の記録環境、作成時間、参照資料、疲労、締切は詳しく報告されていません。一人の人間文書と一回のAI出力を症例ごとに比べたなら、生成者個人の書き方と「人間対AI」が分離できません。

AIの強みを評価するには、人間を不利にする比較でも、AIを不利にする比較でもなく、実務で予定する使い方を比べる必要があります。例えば「人のみ」「AI下書き＋人の確認」「既存テンプレート＋人」を無作為化し、時間、誤り、修正量、本人評価を測る設計です。

4. p値は大きな問いへ答えない

すべてp<0.001という結果は目を引きます。しかし、評価が症例と評価者に繰り返される構造を無視した可能性があり、完全な平均・SD、平均差の信頼区間、標準化効果量も不足しています。

知りたいのは「偶然ではなさそうか」だけではありません。どの程度違い、その差が危険な誤り、時間、信頼、本人の生活へ意味を持つかです。

5. 費用、人員、責任、継続可能性は測られていない

研究はChatGPT-3.5の入力日を報告しましたが、利用料、導入設定、情報管理、教育、確認時間、事故対応、モデル更新の費用を扱っていません。

AI下書きを毎回確認するなら、専門職は最終責任を持ちます。負担が「ゼロから書く」から「もっともらしい誤りを探す」へ移るだけかもしれません。導入前後で、作成時間だけでなく、確認時間、修正率、重大誤り、差し戻し、本人への確認、問い合わせ対応を測る必要があります。

ETIC 2.0への支持、挑戦、更新案

ここからはLeeらが直接検証した結論ではなく、Unknown LabによるETIC 2.0への考察です。

支持する点

研究は、AI出力を専門職の代替ではなく、確認・編集する下書きとして位置づけました。患者・介護者の視点を評価へ含め、質を複数項目へ分けた点も、ケア生態系としてAIを評価する方向と一致します。

AOTAのAI倫理方針も、AIは専門職判断を置き換えず、出力を根拠、正確性、安全性、既存方針との整合で評価するよう求めています[6]。

挑戦する点

ETIC 2.0が「人が最後に確認する」とだけ書けば不十分です。高得点の文書に表の不整合が残るように、人は整った文章を見逃すことがあります。

誰が、どの原情報を、どの順で確認するのか。本人の語りを誰が戻って確かめるのか。誤りが見つかったとき誰が訂正し、共有先へ通知するのか。責任と手順を設計しなければ、人間監督は名目になります。

更新案: AI共同記録の8項目

確認項目	問い
入力と同意	何をAIへ入力し、本人は同意したか
生成範囲	どの文がAI生成で、どこを人が編集したか
事実・推論・提案	三つを文書上で区別できるか
本人の作業目標	本人の言葉、価値、続けたい生活が残っているか
照合と修正	原情報と照合し、何を何分かけて直したか
承認と責任	誰が最終承認し、責任を負うか
停止と訂正	誤りや漏えい時に停止・訂正・通知できるか
成果	時間だけでなく安全、信頼、公平性、作業参加がどう変わったか

これらはETIC 2.0の有効性が実証されたという意味ではありません。Leeらの研究が残した空白を、比較研究で検証可能な設計要件へ変える提案です。

現場でAI下書きを使う前の確認表

- 個人情報を入力してよい契約・環境か。
- 本人へAI利用を説明し、拒否や代替を選べるか。
- 元情報と照合する担当者と時間が確保されているか。
- AIが追加した推論や計画を見分けられるか。
- 本人の言葉と作業目標が、一般的な表現へ置き換わっていないか。
- 修正前のAI出力と最終記録を監査できるか。
- 誤りを他職種や本人へどう訂正するか決まっているか。
- 速さだけでなく、誤り、修正量、本人評価、専門職負担を測るか。

個人情報保護、所属組織の記録規程、契約条件、専門職責任を確認できない環境では、実患者情報を一般向けAIへ入力してはいけません。本稿は特定の製品利用や医療記録の代行を勧めるものではありません。

次に読む論文

次に読むのは、DiMaioらが2026年に発表した「A Language Model for Pediatric Occupational Therapy Documentation」です[7]。

Leeらが架空症例の文書評価を扱ったのに対し、DiMaioらは小児リハ施設で実装しました。10人の療法士がMicrosoft Copilotと、Llama 3 8Bを作業療法記録向けに調整した専用モデルを、それぞれ3週間使いました。

専用モデルの記録は、手書き記録より明瞭性、完全性、関連性、構成で高く評価され、Copilotより簡潔性が高くなりました。しかし、小規模パイロットでは、手作業より記録時間が有意に短くなりませんでした。短いメモを入力したときだけ時間節約が見られ、療法士は次第に詳細な入力へ戻りました[7]。

これは重要です。モデル性能だけでなく、人が安心するためにどれだけ入力し、確認し、修正するかが効率を決めます。Leeらの「高く評価される下書き」から、DiMaioらの「実際に使うと何が起きるか」へ進むことで、AI共同記録の現実的な価値と負担を読めます。

結論

Leeらの15症例比較では、ChatGPT-3.5が作成したA/P記録は、作業療法士が作成した記録より、質と共感の全6項目で高く評価されました。AIは、整理された読みやすい下書きを作る可能性があります。

しかし、これはAIが作業療法士より正しく、共感的で、効率的だという証明ではありません。架空症例、少数評価者、未検証の書面尺度、研究者による文体標準化、入れ子構造を十分扱わない解析という制約があります。本文と表の相関ペアは一致せず、ICCの信頼区間にも誤植があります。

AIの記録を評価するとき、文章の整い方で止まらないことです。

元情報と一致しているか。本人の言葉と作業目標が残っているか。誰が何分かけて直したか。誤りの責任と訂正経路があるか。生活上の成果へつながったか。

共感らしい文章を作ることと、本人の経験へ責任を持って応答することを分ける。そこから、AIと作業療法士の共同記録を設計する必要があります。

記事の位置づけ

本稿は、公開済み査読論文をUnknown Labが批判的に読み解いた未査読の論評記事です。原論文の結果、著者の結論、因果効果の可否、Unknown LabによるETIC 2.0への考察を区別して記載しています。個別の診断、治療、リハビリテーション、医療記録作成、AI製品の利用を指示するものではありません。

執筆: Unknown Lab 編集部（AI支援） / 専門職確認: 作業療法士

参考文献

1 Lee S-A, Park J-H. Artificial intelligence in occupational therapy documentation: Chatbot vs. Occupational Therapists. DIGITAL HEALTH. 2025;11:20552076251386657. doi:10.1177/20552076251386657

2 Ayers JW, Poliak A, Dredze M, et al. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. JAMA Intern Med. 2023;183(6):589-596. doi:10.1001/jamainternmed.2023.1838

3 Değerli Y, Özata Değerli MN, Kars S. Artificial intelligence in occupational therapy: Comparing ChatGPT and occupational therapist approaches in assessment and intervention. Br J Occup Ther. 2026;89(1):6-18. doi:10.1177/03080226251368231

4 Hojat M, DeSantis J, Gonnella JS. Patient Perceptions of Clinician's Empathy: Measurement and Psychometrics. J Patient Exp. 2017;4(2):78-83. doi:10.1177/2374373517699273

5 Mercer SW, Maxwell M, Heaney D, Watt GCM. The consultation and relational empathy (CARE) measure: development and preliminary validation and reliability of an empathy-based consultation process measure. Fam Pract. 2004;21(6):699-705. doi:10.1093/fampra/cmh621

6 American Occupational Therapy Association. Policy E.19: Ethical Use of Artificial Intelligence (AI). Effective April 2025. Official PDF

7 DiMaio R, Tuinstra T, Yu T, et al. A Language Model for Pediatric Occupational Therapy Documentation: Model Development and Pilot Study. JMIR AI. 2026;5:e73274. doi:10.2196/73274

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） / 専門職・統計確認済み