論文とETIC 12

AIの点数は、誰の何を測っているのか

費用予測とケアの必要性から考える、評価指標の妥当性



Daiki Morimoto / Unknown Lab
ORCID: 0009-0009-6444-750X
Version 1.0 / 公開日 2026-10-07
DOI: 10.5281/zenodo.23200634
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
ライセンス: CC BY 4.0

30秒で読む

同じAIの点数を付けられた二人は、同じ程度の支援を必要としているのか。米国の医療データを調べたObermeyerらは、ある費用予測アルゴリズムで同程度のリスクとされた黒人患者の方が、白人患者より多くの慢性疾患を抱えていることを示した。費用を予測する能力と、支援の必要性を捉える能力は同じではなかった。[1]

本稿は、この研究とJacobs・Wallachの測定論を組み合わせ、AIの評価指標が何を意味するかを検討する。[2] 原報の有名な選定シミュレーションには、著者が2022年に公開したコード訂正がある。「対象者中の黒人患者は約47%になる」という推計は59%へ訂正されたが、これは実際の治療成果ではない。[3]

結論は「数値を使うな」ではない。測りたい概念、代わりに数えるもの、選定によって変わる人、実際に生じた結果を分けることである。ETIC 2.0にとっても、本人の選択や関係を重視するという理念だけでは足りない。その理念をどの観察で確かめ、どの結果なら見直すのかを明示する必要がある。

同じ点数なのに、支援の必要性が違う

支援候補者の一覧に、翌年の医療費が高くなると予測された人が並ぶ。その一覧は、複雑な病状を持つ人を見つけ、追加のケアへつなぐために使われる。ところが、頻繁に受診できる人と、仕事や移動の制約で受診を控えている人では、同じ病状でも支払われた費用が異なるかもしれない。後者の支出が少ないことを、必要な支援が少ないことと読むのは適切だろうか。

これは特定の患者の事例を再現したものではなく、今回の研究が明らかにした問題を考えるための場面である。重要なのは、モデルが計算を間違えたとは限らない点にある。将来の支出を正確に予測しても、その支出を「支援の必要性」と読み替える段階で誤りが生じ得る。

似た区別は、医療費以外にもある。相談記録の長さは、相談が十分だったことを意味しない。本人の希望に言及した回数は、その希望が実際の決定に反映された回数ではない。これらは本稿の説明例であり、今回の二論文がすべてを実験したという意味ではない。しかし、測りやすい値を使うことと、その値で大切なものを捉えたことを区別する点では共通する。

前回は、AIを使っている間の成績と、使った後に人へ残る判断の変化を分けた。今回は時間の違いより前に、そもそも成績や変化と呼ぶ数値が何を表しているかを問う。良い対照条件を置いても、比較する双方を不適切な尺度で採点すれば、ケアにとって重要な差を見落とす可能性が残る。

二つの論文をどう組み合わせて読むか

論文	研究の位置付け	主に答える問い
Obermeyerら、2019年、Science	医療データの観察的監査、予測モデル比較、対象選定のシミュレーション	費用を予測する点数は、誰の健康上の必要性を取りこぼすか
Jacobs・Wallach、2021年、FAccT	測定論を計算システムの公平性へ接続する概念・方法論的論文	測りたい概念と観測できる指標を、どの仮定で結んでいるか

表1。主論文の役割。前者の結果を後者の枠組みで読み直す。Jacobs・WallachはObermeyerらの研究を事例として引用しており、独立した二つの臨床的実証として数えることはできない。[1,2]

Obermeyerらの論文は2019年の研究である。古いから無条件に一般化できるわけでも、最近の生成AIだから無関係になるわけでもない。本稿が取り上げるのは、現在の特定製品の性能ではなく、学習目標や評価指標の選択という問題である。当時の商用予測モデルを、現在の対話型AIと同一視しない。

Jacobs・Wallachの論文には、参加者を募集した実験のサンプルサイズや追跡期間はない。教育、司法、医療などの既存事例と測定論を用いて論点

を整理したものであり、介入効果を統合したシステマティックレビューではない。したがって、二論文の結果を一つの効果量へまとめることはしない。

読解では、原著本文に加え、Obermeyerらの著者公開コードと訂正説明を照合した。原著の補足PDFは取得できていないため、補足資料だけで確認できる詳細な定義や解析条件を確認済みとは扱わない。本稿の数値は原著本文・表と著者の明示的な訂正に限定し、原データを独立に再解析したものではない。この資料上の限界も含めて読む必要がある。

研究の対象と、現場での使われ方

Obermeyerらが主解析で扱ったのは、米国のある大規模大学病院で、リスクに基づく契約の対象となっていたプライマリケア患者の2013〜2015年のデータである。黒人患者6,079人、白人患者43,539人、計49,618人を含む。同じ患者を複数年観察しているため、解析上の患者年はそれぞれ11,929、88,080、計100,009となる。100,009人の別々の患者を調べたわけではない。[1]

項目	原報で確認できる内容	読む際の注意
期間・施設	2013〜2015年、米国1大学病院の主解析	現在の日本のケア現場を代表しない
人数・観察単位	49,618人、100,009患者年	人数と患者年を混ぜない
保険	民間保険71.2%、Medicare 28.8%	無保険者を含む全住民の研究ではない
基本属性	平均50.9歳、女性63%	高齢者だけを対象とした研究ではない
自動選定の境界	リスク得点の97パーセンタイル	選定候補になることと実際の参加は別
医師による検討	55パーセンタイルを超える患者を担当医へ提示	医師の判断だけでなく参加資格や運用条件も関わる

表2。対象と運用の条件。人数は本文、患者年と群別属性は原著Table 1を照合した。平均や保険割合は著者の全体記述による。[1]

人種は本人の申告に基づく。黒人群には他の人種・民族的アイデンティティを併せて申告した人も含み、白人群は他の人種・民族を申告していない人とされた。この区分には歴史的・社会的な問いを調べる意義がある一方、複合的なアイデンティティや他の集団の経験を十分に扱えない。二つの群の平均を、各個人の状態と同一視することもできない。

アルゴリズムは、前年の診療・請求データなどから翌年のリスク得点を作る。その得点が、複雑な健康問題を持つ人をケア管理プログラムへつなぐ判断に使われていた。高い得点の患者は自動的に候補となり、より広い範囲の患者は、電子カルテや請求情報とともに担当医へ提示される。人種そのものは予測変数から除かれていた。

ただし、学習の目標は翌年の総医療費であった。「医療費が高くなりそうな人」と「追加の支援を必要とする人」が十分に重なるという仮定が、運用の中に入っていたのである。この仮定は、学習データを増やしたり、同じ費用をより正確に予測したりするだけでは検証されない。

同じリスクで比較すると何が見えたか

著者らは、予測得点を電子カルテの疾患情報や検査値、バイタルサインへ結び付けた。健康状態の一つの指標として、活動性のある慢性疾患の数を用いる。自動選定の境界である97パーセンタイル付近では、黒人患者の慢性疾患数は4.8、白人患者は3.8であり、著者は26.3%多いと報告した。群間差のP値は0.001未満であった。[1]

これは「黒人であることによって疾患が26.3%増えた」という因果効果ではない。同じ予測リスクで比較したときに、二つの群の健康負担が同じではなかったという観察結果である。疾患数だけでなく、血圧や糖尿病関連の指標なども調べており、診断コードの数だけに依存した説明を避けようとしている。

一方、同じ予測得点における費用は、両群で大きくは違わなかった。つまり、費用という予測対象については比較的よく対応しているように見えても、健康状態については対応していない。公平性の評価が、何を正解として採用するかで変わることを示す重要な部分である。

比較の条件	原報値	数値の意味と境界
リスク97パーセンタイル付近	慢性疾患数4.8対3.8、26.3%差、P<0.001	同じ予測リスクでの健康負担差。因果効果ではない
慢性疾患数を一定にする	黒人患者の年間費用が平均1,801米ドル少ない	同程度の疾患数での条件付き差
個別疾患の構成を一定にする	同じく1,144米ドル少ない	調整条件を変えた費用差。上の値と混ぜない
条件付けない群別の平均費用	黒人8,442、白人7,540米ドル	全体平均では黒人群の方が高い。健康負担も異なる

表3。比較条件をそろえて読むための主要数値。原著本文・Table 1・Fig. 1/3に基づく。「黒人患者の費用は常に低い」という要約は誤りである。[1]

表3の最後の行は、とくに重要である。原報の群別平均では黒人患者の医療費の方が高い。それでも、同じ疾患数や疾患構成に条件付けると費用は少ない。これは矛盾ではなく、異なる問いへの答えである。全体平均は各群の疾病負担の違いも含み、条件付き比較はその一部をそろえている。

ここから著者らは、必要なケアと実際に利用できたケアとの間にある隔たりへ注意を向ける。移動、仕事、育児、受診を促す情報、医療者との関係、差別などが候補となる。ただし、このデータから各要因の寄与率を一つずつ確定したわけではない。「この格差は移動手段だけが原因だ」「医療者の偏見が何%を説明した」といった結論は導けない。

17.7%、46.5%、そして訂正後の59%

原報には、リスク得点で選ばれた集団の構成を、病気の多さに照らして入れ替えるシミュレーションがある。自動選定された白人患者のうち比較的健康な人を、境界より下にいる病状の重い黒人患者と入れ替え、選定人数を保ったときに何が変わるかを調べた。実際に患者を除外した介入ではなく、過去のデータ上の推計である。[1]

原報では、選定対象に占める黒人患者の割合が17.7%から46.5%へ増えるとされた。この結果は広く引用されているが、そのまま最新の確認値として紹介することはできない。著者の公開リポジトリには、2022年9月1日付でFig. 1bのコードに関する訂正説明がある。[3]

説明によれば、選定対象から白人患者を一人取り除いた後、次に比較する位置をさらに一つ進めてしまい、次の患者を飛ばしていた。そのため、入れ替えられるべき患者の一部が比較されず、推計を過小評価していた。著者は、訂正後の割合を、原報の約47%ではなく59%と記載している。公開コードでも、削除後に位置を進める処理がコメントアウトされている。

数値・資料	何を示すか	本稿での扱い
17.7%	原報の自動選定集団に占める黒人患者の割合	原報の出発点として記載
46.5%（約47%）	原報の入替えシミュレーション	訂正前の値と明示し、最新値として使わない
59%	著者の2022年コード訂正説明による入替え後の割合	著者報告の訂正値。原データの独立再計算ではない
選定人数を一定にする操作	誰を選ぶかの構成変化	支援を受ける患者の総数や健康改善量ではない

表4。原報と後年のコード訂正。59%に未報告の小数桁や信頼区間を付け加えない。選定集団中の割合を、黒人患者全体の参加率へ読み替えない。[1,3]

訂正は、差の大きさを弱める方向ではなく、強める方向であった。しかし、大きくなった数値をそのまま強い宣伝材料にすることも適切ではない。著者は同じ説明の中で、入替え方式そのものへの批判も紹介している。この方式は、白人患者を外し黒人患者を加える方向に限られ、両群を対称的に再選定するものではない。実務上の変更を最小限にする目的で作った手続きであり、唯一の「公平な配分」を算出したわけではない。

したがって、59%という数値は、費用に基づく選定が大きな取りこぼしを含み得ることを示す材料である。一方で、「公平な医療では黒人患者の割合を必ず59%にすべきだ」という規範でも、「この変更で患者の健康が59%改善する」という効果でもない。訂正の存在と、訂正しても残る方法上の限界を、一緒に扱う必要がある。

予測する目標を変えると、誰が選ばれるか

著者らは、同じ学習方法と予測変数を用い、目的変数を変えた三つの予測モデルも比較した。予測対象は、総医療費、救急受診や入院に関連する回避可能費用、活動性のある慢性疾患数である。データを無作為に約3分の2の学習用と3分の1の検証用に分け、結果は検証用データで示している。ここでも人種は入力から除かれている。[1]

公開コードはLASSOによる実装と合成データでの実行方法を示す。ただし、公開コードを動かした結果を原著の患者データの再現値と扱うことはできない。原データは患者のプライバシーのため公開されておらず、本稿でも実行による独立再解析は行っていない。学習・検証の分離は重要だが、それだけで別施設や将来時点への一般化が保証されるわけでもない。

学習目標	総医療費の集中割合	慢性疾患数の集中割合	選定集団中の黒人患者
総医療費	16.5%（SE 0.3）	10.5%（SE 0.2）	14.1%（SE 0.3）
回避可能費用	14.2%（SE 0.3）	13.0%（SE 0.3）	21.0%（SE 0.3）
活動性のある慢性疾患数	12.1%（SE 0.3）	14.8%（SE 0.3）	26.7%（SE 0.3）

表5。原著Table 2から主要列を再構成。各モデルが高リスクとする上位集団の、検証データ全体に占める費用・疾患数の割合と、その集団内の人種構成。選定境界は97パーセンタイル。SEは標準誤差をパーセントポイントへ換算した値であり、95%信頼区間ではない。[1]

この表で「16.5%」は正答率ではない。総医療費を予測して選んだ高リスク集団に、検証データ全体の医療費の16.5%が集中していたという意味である。「10.5%」も、慢性疾患を持つ人の10.5%を見つけた感度ではなく、疾患数の総量に対する集中割合である。分母を混同すると、モデルの成績について別の話になってしまう。

慢性疾患数を学習目標にすると、選定集団に占める黒人患者の割合は14.1%から26.7%へ変わる。同時に、総医療費の集中割合は16.5%から

12.1%へ下がり、慢性疾患数の集中割合は10.5%から14.8%へ上がる。何を重視するかによって、良くなった指標と下がった指標がある。目的変数を変えれば、すべての性能が一方向に改善するわけではない。

この比較の価値は、同じデータや似た手法を用いても、目標の設定で選ばれる人が変わると示した点にある。ただし、ここで最も病気が多いと予測された人が、追加支援から最大の便益を受けるとは限らない。必要性の大きさ、介入で変えられる余地、本人の希望は別の問題である。

原著は製造企業との協力による、民間保険加入者3,695,943人の全国データの補助的検討も報告する。健康に関する情報を目標へ組み込んだ修正により、同じリスクで比較した黒人患者の過剰な慢性疾患数を集計した指標が48,772から7,758へ減り、約84%の減少とされた。[1] これはバイアスを表す特定の指標の減少であり、黒人患者の病気そのものが84%減ったという治療成績ではない。

人が最後に判断すれば解決するのか

この現場では、アルゴリズムが一人で参加を決めていたわけではない。医師が候補者と周辺情報を見て判断し、参加資格や施設の条件も関わる。著者は、実際の参加と複数の選定ルールのシミュレーションを比較し、人による判断は問題の一部を補っても、目標を変えたモデルほどには構成を変えていないと論じた。[1]

ただし、人の判断が「無意味」と示されたわけではない。シミュレーションに入るのは観察可能な要素に限られ、記録されていない生活事情、本人の希望、利用可能な支援、参加資格の詳細などは十分に表現できない。医師の判断の質を、モデルとの構成比の差だけで評価することはできない。

同時に、人が最後に確認するという手順だけで、指標の問題が自動的に修正されるわけでもない。そもそも誰が候補一覧へ入るか、誰の情報が見やすく提示されるかが、確認の前に決まっているからである。一覧に現れにくい人を、人の善意だけで見つけ続ける運用は脆弱になり得る。

必要なのは、人を減らすか増やすかという二択ではない。何を予測した点数なのかを示し、候補から外れた人も点検できる情報と権限を確保し、その点検にかかる時間と責任を設計することである。この運用提案はUnknown Labの解釈であり、原著が実装効果や必要人員を定量化した結論ではない。

測定の妥当性は、予測精度とは別に問われる

Jacobs・Wallachは、直接観察できない概念を、観察可能な性質から推定する際の仮定を明示する必要を論じる。社会経済的地位を所得だけで表す場合や、教師の働きを試験得点の変化で表す場合を考えると分かりやすい。所得や得点を正確に記録できても、それぞれが概念の全体を十分に表すとは限らない。[2]

このとき、測りたい抽象的なものを構成概念と呼び、それを具体的な変数や手続きへ落とすことを操作化と呼ぶ。測定モデルは両者をどう結ぶかを表し、測定誤差モデルは観察値と誤差についての仮定を表す。予測モデルがどれほど複雑でも、この結び付きが自動的に正しくなるわけではない。

信頼性と妥当性も区別される。同じ状況で繰り返し測って近い値になることは重要だが、一貫して別のものを測っている可能性は残る。評価者が全員、丁寧な文章へ高得点を付けても、その文章が本人の明示した拒否を無視していれば、本人の選択を尊重する能力の尺度としては不十分かもしれない。これは本稿の説明例であり、評価者間一致の高さを否定するものではない。

妥当性の観点	今回の問いへの当てはめ
表面的妥当性	医療費が必要性に関係しそうに見えるだけで十分か
内容的妥当性	受診できない事情や生活上の必要性が指標から落ちていないか
収束的妥当性	別の測定法とも対応するか。同じ偏りを共有していないか
弁別的妥当性	病状の重さと支払能力・受診機会を区別できるか
予測的妥当性	費用以外の関連する結果を、理論に沿って予測できるか
仮説に関する妥当性	測りたい概念について期待される関係が観察されるか
帰結的妥当性	その指標を使うことが、選定や不利益へ何をもたらすか

表6。Jacobs・Wallachが整理する七つの側面を、本稿の問いへ適用したもの。原著の例をそのまま転載したチェックリストではなく、Unknown Labによる応用である。項目を埋めれば妥当性が確定するという意味ではない。[2]

とくに、検証用データで費用予測が当たることと、健康必要性を測る尺度として妥当であることは異なる。学習用に使っていないデータで評価しても、正解ラベルが費用のままであれば、費用との対応を検査したことになる。この点を無視すると、厳密に分割したデータと高度な統計手法が、概念の取り違えを見えにくくする場合がある。

また、公平性には複数の価値判断が含まれる。似た人を似たように扱うこと、集団間の不利益を減らすこと、最も困っている人を優先することは、同じ計算規則ではない。どの集団の割合をそろえたかだけでは、何を公平と考えたかを十分には説明できない。著者らは、数学的な定義の対立として語られる問題の背後に、異なる価値理解があると指摘する。

著者の結論と、そこから言えないこと

Obermeyerらの中心的な結論は、便利で一見有効な代理指標を正解ラベルにする選択が、大きな偏りの源になり得るというものである。費用に関する予測の対応が良くても、健康の必要性について同じ意味を持つとは限らない。目標の変更で選定構成が変わる結果は、この説明に実質的な根拠を与えている。[1]

Jacobs・Wallachの結論は、そのような置換の仮定を測定モデルとして明示し、複数の妥当性の側面から検討することで、公平性の議論を改善できるという方法論的提案である。[2] 七項目を導入した施設で有害事象が何%減るかを検証した論文ではない。

いずれも、ETIC 2.0を直接検証していない。二論文を読んで得られるのは、ETIC 2.0を使った説明にも同じ測定上の要求を課す理由であり、理論の正しさの証明ではない。理念と測定が近いと感じられることと、その理論を用いる方が良い結果を生むことは、別の主張である。

Unknown Labの批判的検討

第一に、費用を疾患数へ置き換えても「本当の必要性」が直接観察できるわけではない。疾患数は診断や受診の機会に影響され、同じ数でも重症度や生活への影響は異なる。検査値を併せて調べたことは強みだが、痛み、役割の喪失、介護負担、本人が望む生活のすべてを表すわけではない。費用という代理指標の問題を見つけたからといって、新しい代理指標が無条件に正解になるわけではない。

第二に、必要性と介入から得られる便益を分ける必要がある。病状の重い人を選ぶことと、ある特定の支援で最も状態が変わる人を選ぶことは一致しない場合がある。さらに、状態を変えにくい人にも支援が必要なことはある。便益の最大化だけで選べば、支え続ける必要のある人を再び外す可能性がある。この選択は、予測精度だけでは決まらない。

第三に、集団間の構成が変わったことと、個々人にとって配分が良くなったことを分けたい。訂正後の59%は重要だが、その値を唯一の到達点にすると、入替え規則や病状指標の選択が見えなくなる。外される人の状態、追加される人の希望、サービスの容量と質を、別に調べる必要がある。

第四に、費用を目標にすること自体を常に誤りとするのは適切ではない。保険財政の見通しを立てる目的なら、費用予測は直接必要な情報となる。問題は、費用を予測した点数を、説明なしに健康必要性やケアの優先順位へ転用することである。同じモデルでも、目的と運用先によって評価基準は変わる。

第五に、追加の測定も負担を生む。本人の希望や生活上の変化を新たに聞き取る場合、誰が説明し、誰が記録し、本人が答えたくないときにどう扱うかが必要になる。評価項目を増やすことを、良いケアと同義にしない。今回の原著は、こうした運用の総費用、必要人員、長期の継続可能性を評価した研究ではない。費用データを使う論文であっても、改善策の費用対効果が確認済みとはいえない。

第六に、著者による訂正の公開を、研究全体の否定とも、すべて解決した証拠とも扱わない。コードの訂正で変わった結果と、変わらない観察所見を区別する必要がある。同じ得点における健康負担差、別の目標を学習させた比較、入替えシミュレーションは異なる解析である。Fig. 1bの訂正だけから、他のすべての数値が同じ割合で変わったと推測してはならない。

最後に、一般化の範囲は狭く保つ。主解析は一つの医療システムの保険加入者であり、全国データでの補助的検討も民間保険加入者を対象とする。日本の作業療法、訪問看護、介護、教育や福祉へ数値を移せない。一方、「過去に使われた資源の量」を「これから必要な支援の量」と同一視していないかという問いは、それぞれの現場で検査可能な問題として持ち帰ることができる。

ETIC 2.0の考え方と重なる点

公開されているETIC 2.0は、技術単体の性能だけでなく、人と環境、関係、活動の中でケアを考える枠組みである。[4] 今回の文献と重なるのは、数値が個人の内側だけから生まれるのではなく、受診できる環境や制度、過去の資源配分に影響されると考える点である。

医療費は、病状の情報を含むと同時に、どのケアへ到達できたかの履歴でもある。その値を本人の必要性だけの表れと読むと、環境によって生じた制約を本人の低い優先順位へ変換してしまう可能性がある。人と環境の関係を確認する姿勢は、この変換に気付く手掛かりになる。

ただし、この問題は構成概念妥当性や代理指標の誤指定という既存の測定論で十分に説明できる可能性がある。ETIC 2.0という名称を加えること自体に、独自の予測力や改善効果があるとは示されていない。既存理論で説明できる部分を、独自理論の成果として取り込まないことが重要である。

答えるべき問い

ETIC 2.0を用いて「本人中心」「関係が良い」「選択を支えた」と評価するとき、それぞれを何によって観察するのか。丁寧な説明、本人の満足、選択肢の提示、実際の選択、選択後の生活は互いに関係するが同じではない。どれを測ったのかが曖昧なままでは、良い言葉を並べるほど高得点になる尺度を作ってしまう。

また、誰の結果を改善と呼ぶのか。専門職の作業時間が減った一方で家族の確認負担が増えた場合、本人の選択肢が増えた一方で説明が難しくなった場合、合計点だけで評価してよいのか。主体別に結果を示すことと、全員の利益が常に一致すると仮定することは違う。

さらに、例外を見つけるための評価が、別の不利益を生まないか。本人の語りを毎回詳しく集めることは、望まない情報開示や回答負担を増やし得る。重要なものを測りたいという理由だけで、測定する権限が自動的に生

まれるわけではない。目的、最小限の情報、拒否できる条件を伴わせる必要がある。

具体的な見直し案

以下はUnknown Labによる未検証の提案であり、原著の臨床的推奨ではない。新しい大規模な総合尺度を作る前に、既存の評価項目の意味を小さく点検する。対象を広げるより、各項目で言える範囲を狭く正確にする方を優先する。

評価したいこと	観察する候補	取り違えを見つける例
本人の選択を尊重する	明示した希望・拒否が提案と決定に反映されたか	文体は丁寧だが本人の拒否を無視する回答
必要な支援へつなぐ	未充足の必要性、接続の実行、取り残された人	利用回数は少ないが移動や費用の制約が大きい人
安全に判断する	重大な見逃し、適切な保留・相談	総合点は高いが一つの危険な勧告を含む回答
負担を減らす	本人・家族・専門職それぞれの時間と負担	専門職の時間短縮が家族の無償作業へ移る場面

表7。評価指標を点検するための提案。実患者に使用する検証済み尺度ではなく、研究設計上の例である。表現上の好ましさと実際の選択・接続・安全・負担を分ける。

比較研究では、同じ情報、時間、利用可能な資源を与えた既存の構造化手順を対照にする。独自の枠組みを使う条件だけに丁寧な入力や長い検討時間を与えない。評価者に条件名を見せず、事前に定めた項目で判断し、設計中に何度も使った例とは別の課題でも確かめる。この基本条件がなければ、理論の価値と作業量の差が混ざる。

反例は、評価尺度の弱点を探すために使う。例えば、丁寧な文章と本人の選択の尊重を意図的に食い違わせ、その尺度が違いを検出できるかを調べる。これは現場で危険な対応を試すという提案ではなく、研究用の記述課

題で尺度の意味を点検する段階の話である。実患者への研究には別の倫理的・専門的検討が必要となる。

結果は、総合得点、重大な失敗、主体別の負担を分けて示す。平均が上がったから重大失敗も減った、本人の満足が上がったから長期的な選択可能性も増えた、と推測で埋めない。費用や時間の増加も併記し、追加測定を続ける価値があるかを判断する。

何が分かれば、提案を小さくできるか

もし既存の短い尺度と汎用の構造化手順で、本人の選択、重大失敗、主体別負担を十分に捉えられ、追加項目が新しい取りこぼしを発見しないなら、新たな概念や評価票を増やす必要はない。測定の意味を明記するだけで足りる可能性もある。

逆に、総合点の高い回答ほど本人の拒否を無視する、利用の少ない人ほど必要な支援から外れる、といった事前に想定した不一致が繰り返し確認されれば、現在の尺度への信頼は弱まる。その場合も、原因を一つの理論へ帰属させず、採点基準、情報不足、課題の偏り、運用上の制約を比較して検討する必要がある。

有意差がないことだけで同等と結論しないことも重要である。許容する見落としの幅や重大失敗の基準を事前に決め、十分な不確実性の評価を伴わせる。今回の二論文は、この比較に必要な万能の閾値やサンプルサイズを与えてはいない。提案を反証に開くとは、どの結果なら縮小・保留するかを残すことであり、検証前から複雑な体系へ固定することではない。

次に読む論文

Piersonらの「An algorithmic approach to reducing unexplained pain disparities in underserved populations」を推薦する。[5] 膝のX線画像から、従来の画像上の重症度だけでなく、患者が経験する痛みを予測する方向へ学習目標を置いた研究である。今回の費用と必要性の問題から、本人の報告を目標へ近づけると何が変わるかへ問いを進められる。

次に確認したいのは、患者の報告を用いれば自動的に正しい尺度になるのか、どの人の経験が学習データに含まれる必要があるのか、治療の候補判定と治療後の便益をどう分けているのかである。本稿では次読候補として紹介するにとどめ、その定量的結果を今回の論証へ加えない。Obermeyer、Mullainathanが共著者であり、完全に独立した研究チームの追試という位置付けでもない。

結論

AIが何かを正確に予測したとしても、それが現場で知りたいものと同じとは限らない。費用をよく予測する点数が、健康上の必要性を等しく表しているとはいえないという研究は、この区別を具体的に示した。後年のコード訂正を含めて読むと、目立つ一つの数値だけで結論を作る危うさも見えてくる。

必要なのは、数値への不信でも、測れないものの無条件な称揚でもない。何を測りたいのか、代わりに何を数えたのか、誰の経験が落ちるのか、その指標を使って何が変わったのかを分けることである。ETIC 2.0も、この問いから免れるものではない。

本稿は公開研究に基づく未査読の批判的論評であり、新たな臨床試験、原データの独立再解析、ETIC 2.0の有効性の証明ではない。良い評価とは、得点を増やすことだけでなく、その得点ではまだ言えないことを残しておくことでもある。

参考文献

1. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366:447-453. https://doi.org/10.1126/science.aax2342（著者公開の原文全文・本文表を参照。出版社補足PDFは未取得）

2. Jacobs AZ, Wallach H. Measurement and Fairness. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. 2021:375-385. https://doi.org/10.1145/3442188.3445901（著者公開全文: https://arxiv.org/abs/1912.05511 、v3を参照）

3. Obermeyer研究チーム. Public Release Synthetic Data and Code: Dissecting Bias. README「Error in code that makes Figure 1b」、2022-09-01訂正説明、code/figure1/figure1b.R. https://gitlab.com/labsysmed/dissecting-bias（2026-10-07確認。59%は著者の訂正報告値）

4. Morimoto D. エコテクノロジカル・インテグラティブ・ケア(ETIC)2.0, Version 2.1改訂版: AI時代の作業療法とケア生態系設計. Unknown Lab; 2026. https://doi.org/10.5281/zenodo.21848456

5. Pierson E, Cutler DM, Leskovec J, Mullainathan S, Obermeyer Z. An algorithmic approach to reducing unexplained pain disparities in underserved populations. Nature Medicine. 2021;27:136-140. https://doi.org/10.1038/s41591-020-01192-7（次に読む論文。今回の主解析対象ではない）