論文とETIC 14

AIが答えを保留すれば、安全性は上がるのか

回答の精度と、引継ぎ後の結果を分ける



Daiki Morimoto / Unknown Lab
ORCID: 0009-0009-6444-750X
Version 1.0 / 公開日 2026-10-21
DOI: 10.5281/zenodo.23520819
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

AIが自信のない回答を控えると、残った回答の正確さは上がりやすい。その効果を安全性の面から評価するには、保留後の担当者、待ち時間、未解決の依頼まで追う必要がある。

Jonesらは、信頼度の低い予測を保留すると全体の平均正答率が上がっても、もともと成績の悪い集団では改善が小さい、あるいは悪化する場合があると示した。[1] Swaminathanらは臨床記録の分類で、曖昧な記録を人へ戻す仕組みを検討した。ただし、精度の数値は保留した記録を除いた部分の評価であり、実際の患者の安全性や人員削減を検証した試験ではない。[2]

本稿では、回答部分の精度、全依頼に対する回答率、人への引継ぎ結果を評価の軸に据える。ETIC 2.0にも、保留によって生じる本人や支援者の負担をどう捉えるかという課題がある。

1. 「担当者に確認してください」の先は、誰が引き受けるか

相談窓口にAIを導入したとする。確信を持てない質問には「担当者に確認してください」と返す。回答の誤りは減ったが、担当者の机には確認待ちが積み上がり、相談した人は返事を待ちながら同じ説明を繰り返している。AIの成績表には、この待ち時間も手間も載っていない。

これは説明用の仮定である。情報不足の際に断定を避け、専門職へつなぐ対応は妥当だろう。ただし、受入れ先や返答の期限が決まっていなければ、利用者の問題は残る。

顧客対応、学校の相談、職場の承認、行政手続きでも、例外的な依頼を人へ戻す設計は考えられる。適切な確認によって誤った決定を防げたのか、それとも処理が滞ったのか。引継ぎ後の結果を記録すれば、運用を改善する手掛かりになる。

前回は専用AIと汎用AIの比較を扱い、拒否した回答を除いた平均点の限界を指摘した。今回は、保留された依頼も含めて全体の成績を評価する方法を検討する。

2. 二つの論文が扱うもの、扱わないもの

論文	研究の対象と方法	本稿で担う役割
Jonesら、ICLR 2021	画像・文章の5データセット、選択的分類の実験と理論解析	平均精度の改善が集団間の格差を隠す仕組み
Swaminathanら、JAMIA 2024	過去の臨床記録を分類する後ろ向き研究、二つの閾値と人の確認	保留を含む処理を、誤りと確認負担の費用で評価する試み

表1。二論文の役割。JAMIA論文のオンライン公表は2023年9月であり、巻号は2024年である。どちらも現在の対話型AIを患者へ割り付けた臨床試験ではない。[1,2]

選択的分類とは、一定の条件で予測を保留する方法である。一般にはモデルが計算した信頼度を使い、低い入力を保留する。高い信頼度を示す誤答もあるため、回答を選別した後の精度を確かめなければならない。

Jonesらは回答の質を集団別に調べ、Swaminathanらは誤分類の損失と人が確認する負担を比較した。人への引継ぎによって集団間の格差を解消できるかは、両研究を通じて残る問いである。

Jonesらの公開原文と全付録、Swaminathanらの出版社HTML全文と補足資料を照合した。患者単位の原データは再解析していない。後者の資料間で一致しない数値や計算は、該当箇所で明記する。

3. 正答率が上がっても、答えられた人は減る

選択的分類では、回答した入力のうち正しかった割合と、全入力のうち回答した割合を併記する。後者はcoverageと呼ばれ、本稿では回答率と訳す。保留率は、その補数に当たる。

仮に100件の依頼があったとする。全件に答えて80件が正しかったモデルと、50件に答えて45件が正しかったモデルでは、回答部分の正答率は80%と90%になる。後者の回答率は50%であり、残る50件の対応も含めて有用性を判断する必要がある。

保留には、有害な依頼の拒否や必要な確認への接続も含まれる。すべてを誤りとして扱うと、危険な断定を促しかねない。一方、回答部分の正答率だけを高めようとすれば、簡単な依頼に回答を絞れる。両指標を併記するのは、こうした偏りを見つけるためである。

観察するもの	分母	分からないまま残ること
回答部分の正答率	実際に回答した依頼	保留された依頼の行方
回答率・保留率	受け付けた全依頼	回答や保留が適切だったか
集団別の正答率と回答率	それぞれの集団	集団内の個々の不利益
引継ぎ後の完了・未解決	保留して人へ戻した依頼	追跡できなかった人の結果
全体の重大失敗と負担	事前に定義した全依頼	測定外の長期的影響

表2。本稿による評価の整理。最後の二行は、二論文で十分に測定されていない後続評価の候補である。

閾値を動かして回答率と正答率の関係を描くと、どれだけの入力を保留することで、どの程度の精度を得るかが見える。この曲線は、一つの閾値で報告する高い精度より情報が多い。ただし、どの位置を運用に選ぶかは曲線だけでは決まらない。誤りの重大さ、人の確認能力、待てる時間、本人の選択が関わるからである。

4. 平均を良くする選別が、ある集団を悪くする

Jonesらは、信頼度で回答を選別する標準的な方法を、画像と文章の5データセットで調べた。課題は髪の色の分類、文章の有害性、画像の種類、胸部X線所見、文同士の意味関係である。各課題に個別の学習・評価用データを用いた。[1]

著者らが注目したのは、正解と本来無関係な手掛かりにモデルが依存する状況だった。例えば、胸部X線で胸水を分類するとき、支持器具の有無へ頼り過ぎれば、胸水があるのに器具がない例で失敗し得る。このデータでは相関を強める訓練標本を作っており、その条件を自然の臨床頻度と取り違えてはいけない。

群は、正解ラベルと、疑わしい手掛かりとなる属性の組合せで定義した。「最も成績の悪い群」は全件に回答した時点で選び、閾値を変えた後も同じ群を追跡した。

主な方法は、モデルの出力確率から信頼度を得て、低いものを保留する方法だった。付録では、計算の一部を変えて繰り返し予測した際のばらつきを利用する方法でも確認している。結果は、全体平均では回答を絞るほど正答率が高まっても、低成績の群では改善が小さかったり、逆に下がったりするというものだった。[1]

ある群で誤答の信頼度が高い場合、選別によって正答が先に除かれ、誤答が残ることがある。その群では、回答を絞るほど正答率が下がり得る。

全体の回答率を100%から50%へ下げたMultiNLIの例では、最も低成績の群の正答率は65%から75%へ上がった一方、次に低い群は77%から95%へ上がった。前者は10、後者は18パーセントポイントの改善である。[1] 両群で改善したものの、格差は広がった。なお、50%は全体の回答率であり、各群の回答率は個別に確認する必要がある。

群ごとの損失へ配慮する訓練では、格差を軽減できる可能性が示された。ただし、訓練時に群の情報を必要とし、未知の不利益集団への適用には課

題が残る。低回答率では例数が少なく、推定も不安定になる。理論解析の結論にも、分布に関する仮定が付く。

この研究は、保留による精度改善に集団差が生じることを示した。人へ戻した後の費用や、群によって回答率が異なることの公平性は、今後の課題とされている。

5. 臨床記録を「判定する・保留する」に分ける

Swaminathanらは、診療記録から病名や手術の情報を抽出するモデルを検討した。研究用の対象者を選ぶ際には、目的の診断や処置を記録から確認する。研究の狙いは、人が読む負担と、患者を誤って含める・除外する損失の調整にあった。[2]

資料はMIMIC-IIIとStanfordの電子診療記録に由来する。前者の元データ期間は2001〜2012年、後者は1998〜2022年であり、直腸がんの候補選定は2012年1月〜2021年12月だった。これらは過去の記録の収集期間を示す。AI導入後の追跡調査は行われていない。

うつに関する課題では、人が注釈を付けた1,138件から、不一致や不確実な注釈を含むものを除き、563人の563記録を残した。うち136件が陽性であった。112件を開発、451件をテストに用い、別の未注釈1,000件には語句などの規則で弱いラベルを付けた。「1,563件」という数を、すべて専門家が確定した独立した患者の正解データと紹介するのは不正確である。

この除外によって、評価対象は人の判断が一致した記録に限られる。現場で判断に迷う記録への対応を確かめるには、除外例を含む追加評価が必要となる。

課題	患者選定と記録数	訓練 / テストの記録数
うつの記載	注釈1,138件から563人・563件、別に弱ラベル1,000件	開発112、テスト451、弱ラベル訓練1,000
神経膠芽腫（GBM）診断	候補1,195人中629人を抽出、659件	395 / 264
直腸腺がん（DRA）診断	候補721人中298人を無作為抽出、601件	361 / 240
腹会陰式直腸切断術（APR）	同じ直腸がん患者群、986件	592 / 394
低位前方切除術（LAR）	同じ直腸がん患者群、882件	529 / 353

表3。本文と補足Table S2による対象。患者数と記録数を区別する。APR/LARの件数は本文に名称の重複があるため補足表を採用した。各行を足して独立した患者総数を作らない。[2,3]

GBMでは、病理報告に該当語を含む成人を候補とし、原発性の診断を対象にした。再発や別の病変からの転化は同じ陽性として扱わなかった。直腸がんでは診断・処置コードで候補を選び、年齢や病理の条件で絞った。したがって、病名の手掛かりがない人を含む全住民からのスクリーニング成績ではない。

Stanfordの課題では、記録を無作為に60%と40%へ分け、前者で訓練、後者で評価した。同じ患者が複数の記録を持ち得る一方、分割は記録単位である。患者が訓練側と評価側に完全に分離されたとは確認できない。似た表現や経過が両側へ入る可能性を考えると、未知の患者や別施設での成績には追加検証が必要となる。

6. 二つの閾値と、人に確認する費用

臨床記録モデルは、語や連続する語句の出現を特徴量にし、正則化したロジスティック回帰で分類した。多数の変数を使う場合の複雑さを抑え、交差検証で設定を選んでいる。

通常の二値分類では、一つの閾値より上なら陽性、下なら陰性とする。研究では二つの閾値を置き、低い閾値より下は陰性、高い閾値より上は陽性、その間は未確定として人が確認する対象にした。[2]

閾値は、偽陽性、偽陰性、保留の相対的な損失から決める。ここでいう費用は金額ではなく、望ましくない結果への重みである。誤分類を一件避けるために何件の読み直しを受け入れるか、という選好を使う。

著者らは臨床家と記録抽出担当者から7件の回答を得て、一つの外れ値を除き、6件の平均を使った。偽陽性、偽陰性、保留の重みは13.5対8対1となった。回答者が何件の確認を受け入れるかには幅があり、偽陽性に対して10〜20件、偽陰性に対して3〜15件だった。少人数の選好を、すべての施設や患者が共有する価値判断と見なすべきではない。[2]

調査では、人が確認すれば確実に正しく決定できると仮定している。現実の確認精度は経験や記録の曖昧さに左右され、必要な時間も変わる。保留一件の費用を一定の1とする設定は、こうした差を単純化している。

保留に伴う確認負担を評価へ組み込んだ点は、この研究の強みである。相談支援へ応用するなら、本人や家族の待ち時間、誤りの重大さも含めて重みを検討する必要がある。

7. 自動判定した割合と、残った記録の精度

GBMのテスト対象264件のうち、モデルが自動判定したのは150件で、114件は保留された。自動判定の割合は約57%である。表にある感度0.96と特異度0.96は、この保留を除いた部分の指標であり、264件を人とAIで処理した後の最終的な感度・特異度ではない。[2]

課題	全記録 / 自動判定 / 保留	感度: 単閾値 → 二閾値	特異度: 単閾値 → 二閾値
GBM	264 / 150 / 114	0.94 → 0.96	0.79 → 0.96
DRA	240 / 218 / 22	0.87 → 0.93	0.92 → 0.93
APR	394 / 375 / 19	0.82 → 0.91	0.99 → 0.99
LAR	353 / 318 / 35	0.71 → 0.91	0.99 → 1.00

表4。原著Table 2の主要値を再構成。保留件数は全記録数から自動判定数を引いて計算した。二閾値の精度は、保留後に残った記録で評価されている。APRの単閾値感度は本文0.81と表0.82が異なり、ここでは表値を採用した。[2]

感度は陽性の記録をどれだけ陽性と判定したか、特異度は陰性の記録をどれだけ陰性と判定したかを表す。保留例を除くと評価対象の構成が変わるため、表の矢印には対象を選別した影響も含まれる。

保留の偏りも無視できない。原著Table 2では、APRの陽性記録の保留率は25%、陰性記録では2.3%、LARではそれぞれ32%と4.9%だった。[2] 全体の保留率が低くても、見つけたい出来事のある記録を、人へ多く戻す設計になり得る。これは患者属性間の公平性を調べた結果ではないが、全体平均だけでは残る仕事の構成が分からないことを示す。

著者らは自動化できる割合を57〜95%と表現する。しかし、それは記録の割合であり、職員の労働時間が57〜95%減ったという測定ではない。保留された少数の記録ほど難しければ、読む時間の多くが残ることもあり得る。確認準備、誤りの点検、連絡、記録の修正といった作業も、記録数の補数だけでは見積もれない。

うつの分類では、保留が増えても費用が下がらない条件があった。偽陰性の重み5、偽陽性2、保留1という設定のロジスティック回帰では、保留率37.25%で、費用は267から290へ約9%増えた。別モデルでは同じ設定で費用が減っており、利益はモデルにも依存した。[2,3]

8. 原報の数値に残る不一致

原著は、がん関連4課題で費用が16〜51%減ったと報告している。ただし、本文の費用式、示された重み、Table 2の件数と費用値を照合すると、一部を同じ手順で再計算できなかった。これは確認可能な資料上の留保であり、研究不正や研究全体の誤りを意味するという指摘ではない。

GBMでは、偽陽性2件、偽陰性4件、保留114件に13.5、8、1を掛けると合計173となり、掲載値に一致する。一方、DRAで同じ計算を行うと186となり、掲載値114とは一致しない。本文の式には総観察数で割る記載もある。これらの違いが、別の費用設定、表記、集計段階の差のどれによるかは、公開資料から確定できない。[2,3]

確認箇所	資料間の違い	本稿での扱い
APR/LARの記録数	本文で術式名が重複、補足S2ではAPR986・LAR882	補足表に従い、差を明記
APRの単閾値感度	結果文章0.81、Table 2は0.82	表値を採用し、勝手に統一しない
費用値	GBMは再計算一致、直腸がん関連は示された重みでは一致しない	検証済みの費用削減率として採用しない
精度の分母	二閾値では保留を除外	全体の安全性へ拡張しない

表5。本文と補足資料を照合した本稿の点検結果。公式な訂正文は未確認であり、不一致の原因は未解決として扱う。

本稿では、照合できた自動判定・保留件数と精度を主要数値として採用した。費用削減率の確認には、原データの再解析や著者への照会が必要となる。

誤りの点検では、人が付けた正解ラベルの修正、再訓練、指標の再計算も報告されている。こうした改善過程で評価データを繰り返し参照した場合には、修正に使っていない独立した凍結テストで最終成績を確かめたい。

9. 著者の結論と、観察から言える範囲

Jonesらは、選択的分類の運用に際し、平均精度に隠れた群別の悪化を点検する必要があると結論した。[1] 群情報を使う訓練の有用性も示されたが、適用条件の検証は残る。

Swaminathanらは、誤分類より人の確認の費用が小さい場面で、選択的予測が臨床記録の抽出を支援し得ると結論した。[2] 実際の確認完了までを追う費用対効果や、患者への有害事象は評価していない。

二論文の成績差は、既存データ上で計算条件や分類規則を変えて得られた結果である。導入による健康改善や事故減少の因果効果を確かめるには、現場での比較研究が必要となる。

対象と課題が異なるため、両研究の結果は個別に解釈する。本稿では、群格差と確認負担という二つの観点から、保留の評価に不足する項目を検討した。

10. Unknown Labの批判的検討

保留の目的に応じた評価が必要である。情報不足には追加質問、専門判断には適切な相手への委譲、有害な依頼には拒否が求められる。各対応がその目的を果たしたかを判定したい。

引継ぎでは、受け手による受入れの確認が欠かせない。「専門家に相談を」と案内しても、相談先の不在、連絡手段の不足、受付時間外といった事情で接続できない場合がある。委譲した側のログに、受入れ側の状況を加える必要がある。

人の確認精度について、JAMIA研究はGBMの172記録を重複抽出し、うち54記録を医師とも比較した。評価者間一致の指標κは0.869、非医師の感度は0.98だった。[2] 確認手順を点検した点は評価できる。運用時には、難しい保留例への対応力と処理可能な件数も調べる必要がある。

緊急性のある相談を通常の確認待ちへ回すなど、重大な失敗は独立して集計すべきである。平均正答率では少数の深刻な見逃しが埋もれやすい。重大さの基準は、専門職や当事者を含めて運用前に決める。

群別の点検では、稀な状態や複数の困難を持つ人など、既知の属性だけでは捉えにくい不利益にも留意する。ただし、細分化で例数が減ると推定は不安定になる。各群の分母と不確実性、事前に定めた分析か探索的な分析かを併記する必要がある。

負担の評価には、家族による説明の繰返し、専門職による根拠確認、本人の待ち時間も含めたい。誰に負担が集中したかを示すことで、総費用だけでは見えない問題を捉えられる。測定自体の負担も考慮し、重大失敗と未解決を把握できる最小限の記録を優先する。

今回の問題は、既存の選択的予測と人への委譲の理論で十分に説明できる可能性がある。新しい枠組みの独自性を論じるには、既存の方法を適切に実装した対照との比較が必要となる。

11. ETIC 2.0の考え方と重なる点

ETIC 2.0は、人、環境、関係、活動の中でケアを捉える。[4] 保留した依頼の受け手や、次の行動を可能にする条件へ目を向ける点で、今回の論点と重なる。

保留は、相談者の待ち時間、担当者の仕事、組織の責任配分に影響する。受入れ可能な相談先と返答の期限を整えることは、ケアの環境を設計する具体的な課題となる。

ETIC 2.0の有効性は、両論文では検証されていない。一般的な引継ぎ手順で十分な改善が得られる可能性を含め、追加価値を確かめる必要がある。

12. 答えるべき問い

「適切な保留」の判定基準は、必要な情報を尋ねたか、妥当な相談先を示したか、接続まで確認したかといった行動で定めたい。慎重な言葉遣いだけで高得点を与えると、実際の対応を評価し損ねる。

未解決の依頼を誰が引き受けるかも決める必要がある。再入力、担当者への照会、別部署への移管、期限付きの再評価について、必要な資源と責任者を明らかにする。

比較条件としては、回答率をそろえて精度を比べる方法や、重大失敗の許容範囲をそろえて回答率を比べる方法がある。閾値を変えたときの結果も示せば、特定の設定だけを有利に選ぶ危険を減らせる。

追跡できなかった依頼は未確認として残し、接続失敗や未解決も記録する。返答のない依頼を解決済みと扱えば、支援へ到達できなかった人を見落とす。ただし、追跡には本人の意思と個人情報の取扱いに関する配慮が必要である。

13. 具体的な見直し案

以下はUnknown Labによる未検証の研究設計案である。まず、個人情報を含まない研究用の記述課題で、同じ基礎モデル、同じ入力情報、同程度の資源を用い、一般的な構造化手順との比較から始める。臨床利用の適否は、この段階では判断できない。

記録する段階	最小限の観察項目	避けたい取り違え
依頼の受付	全件数、事前に定めた課題・集団	保留例を最初から解析対象外にする
AIの応答	回答、追加質問、委譲、拒否と理由	全ての非回答を一律に加点・減点する
回答部分	正確さ、重大失敗、回答率	簡単な例だけ残した精度を全体の改善と呼ぶ
引継ぎ	受入れ可否、担当、期限、必要情報	案内文の存在を接続完了と見なす
結果	完了、未解決、未確認、主体別負担	返答のなさを問題解決と見なす

表6。全依頼を追うための最小限の設計例。実患者を対象としたデータ収集には、別の倫理的検討と専門職・当事者確認が必要である。

両条件には、同じ人の受入れ容量と相談先を与える。一方だけが無制限に専門家へ戻せると、成績差に外部資源の差が混ざる。比較対象にも十分に良い保留・確認の指示を与え、既存の方法で残る重大失敗を減らせるかを検討する。

評価者には条件名を隠し、事前に決めた基準で、適切な回答、必要な追加質問、妥当な委譲、不適切な拒否を判定する。最終評価には設計中に使っていない課題を用いる。判断が割れた例は保存し、評価基準の曖昧さを点検する。

記述課題で評価できるのは、提示された条件での引継ぎの妥当性までである。実際の待ち時間、確認負担、引継ぎ完了率は、現場での検証課題として残る。

14. どの結果なら、提案を縮小するか

観察された結果	見直すべき判断
回答部分の精度だけが上がり、全依頼の未解決が増える	安全性や有用性が向上したという主張を保留する
同じ平均精度でも、特定群の重大失敗・保留が偏る	平均だけで運用を採用せず群別条件を再検討する
一般的な構造化対照で同等の結果が得られる	独自機構の追加価値を主張せず設計を簡素化する
人への引継ぎを考慮すると優位性が消える	AI単体の評価と運用全体の評価を分離して報告する
適切な保留が重大失敗を減らし、負担も許容範囲に収まる	その条件に限った次段階の検証を検討する

表7。本稿の提案を検証・縮小するための条件。良好な結果が得られても、対象・時期・運用条件を変えた検証が必要となる。

同等性を評価するには、許容できる差と必要な例数を事前に定める必要がある。有意差の不在だけでは同等性を示せない。重大失敗の定義や不確実性の表示も含め、今回の比較課題に即して設計する。

既存の手順で同等の結果が得られるなら、枠組みを追加せず簡素な設計を採用する。保留によって利益が得られる条件を絞り込むことも、この比較の成果となる。

15. 次に読む論文

Presacanらの「When silence is safer: a review and decision-theoretic framework for LLM abstention in healthcare」を次読論文として挙げる。[5] 不確実だから回答を控えることと、有害な行為を助けないために断ることを分け、医療対話での保留を評価する方向へ進むための文献である。

次に確認したいのは、追加質問、専門家への委譲、拒否の判定基準と、人の受入れ容量や未解決の相談を扱う範囲である。本稿では次読候補として紹介し、定量的な結論の根拠には用いていない。

結論

保留は危険な断定を避ける選択肢となるが、安全性への寄与は、その後の対応に左右される。今回の研究は、回答部分の精度が改善しても集団間の格差が広がり得ること、人の確認負担を含めると利益が条件によって変わることを示している。

運用を評価する際は、全依頼を対象に、回答の精度、保留された人の偏り、引継ぎの成否、未解決と負担を確認したい。担当者と返答の期限が定まり、依頼が実際に解決したかまで追うことが、改善の手掛かりになる。

参考文献

1. Jones E, Sagawa S, Koh PW, Kumar A, Liang P. Selective Classification Can Magnify Disparities Across Groups. ICLR. 2021. https://doi.org/10.48550/arXiv.2010.14134（著者公開原文v3、2021-04-14、本文・全付録を参照: https://arxiv.org/abs/2010.14134 ）

2. Swaminathan A, Lopez I, Wang W, et al. Selective prediction for extracting unstructured clinical data. Journal of the American Medical Informatics Association. 2024;31(1):188-197. https://doi.org/10.1093/jamia/ocad182（オンライン公表 2023-09-28。出版社HTML全文を参照）

3. Swaminathan A, et al. Supplementary information for ocad182. Tables S1-S6 and Figure S1. https://academic.oup.com/jamia/article/31/1/188/7285661#supplementary-data（同論文の補足DOCX、2026-10-21確認）

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. Presacan O, Nik A, Ojha J, Thambawita V, Ionescu B, Riegler MA. When silence is safer: a review and decision-theoretic framework for LLM abstention in healthcare. npj Digital Medicine. 2026;9:749. https://doi.org/10.1038/s41746-026-02882-1（次に読む論文。今回の主解析対象ではない）