厚生労働科学研究費補助金(長寿科学政策研究事業)
「機械学習を用いた介護認定審査会の審査判定プロセス等を補助するシステムを開発するための研究」
研 究 報 告 書

医療・公衆衛生における人工知能(AI)を用いた統計解析に関する予備的文献検討

研究代表者　高橋 秀人（帝京平成大学 薬学部）

研究要旨

AI 技術が医療・公衆衛生の統計解析に広く取り込まれつつあるが，その手法・応用・課題・報告基準の適用状況を領域横断的に整理した総括的レビューは， 現状のところ見当たらない．本稿は，この領域の現状を把握するための予備的なレビューである．主要トピックについてウェブ検索エンジンによる検索を計 7 回行い，その結果から各領域を代表する文献を選択的に収集して 34 件に整理した.これらを報告・評価基準，予測モデルの方法論，因果推論，自然言語処理・LLM，サーベイランス，公平性・説明可能性の 6 領域に整理した．予測モデルの領域では，報告・方法論的品質の低さが複数の系統的レビューにより指摘されており，ある系統的レビューでは TRIPOD 項目の遵守が中央値で約 4 割，キャリブレーションと外的検証の報告が乏しいと報告されている．われわれは，報告の透明性・キャリブレーション・外的検証・公平性・実装の検証という共通の課題について，推定対象の定義・仮定の開示・適切な検証・誠実な報告という従来の統計的推論の方向性が AI にも必要と考える．

A. 研究目的

機械学習や深層学習，そして近年の大規模言語モデル(LLM)は，医療・公衆衛生のデータ解析に次々と取り込まれている．われわれがまず注目したのは，これらの応用が予測，因果推論，自然言語処理，感染症サーベイランスと広範に及ぶにもかかわらず，それらを一つの視座から整理した総括的なレビューが見当たらないという点である．個別領域のレビュー ― 例えば LLM の臨床応用，感染症サーベイランスへの AI の応用，因果推論への機械学習の導入 ― は存在するが，領域を横断して手法・目的・課題・報告基準の適用状況を俯瞰したものは，われわれの知る限り見当たらない．

本稿は，この空白を念頭に，医療・公衆衛生における AI 統計解析の現状を把握するために実施した予備的なレビューである．問うのは「AI が有効か」という効果の検証ではなく，「どの手法がどの目的に用いられ，どこに課題があるか」という見取り図的なものである．

B. 研究方法

本稿で取り上げた文献は，次の手順で予備的に収集したものである．医療・公衆衛生における AI 統計解析の主要トピックについて，ウェブ検索エンジンによる検索を計 7 回行った．検索語は，(1) TRIPOD+AI(予測モデル報告ガイドライン)，(2) PROBAST-AI(バイアスリスク・機械学習)，(3) 因果推論・機械学習・疫学(ターゲット学習)，(4) 大規模言語モデル・公衆衛生，(5) アルゴリズム公平性・説明可能性・臨床 AI，(6) AI・感染症サーベイランス・アウトブレイク予測，(7) 機械学習・予測モデル・キャリブレーション・外的検証・報告，である．各検索が返した上位の結果から，各領域を代表すると判断した文献を選択的に収集し，報告・評価

基準，予測モデルの方法論，因果推論，自然言語処理・LLM，サーベイランス，公平性・説明可能性の 6 領域に整理して，計 34 件とした．

検索に用いたのは学術データベースではなく一般のウェブ検索エンジンであり，文献の収集は悉皆的でなく著者の選択に依存する．したがって，本稿の 34 件および領域別の整理は予備的な把握であって，件数や分布を網羅的・定量的な結果として解釈すべきものではない．データのチャート化は，書誌情報・解析目的の領域・AI 手法やトピック・主な知見と課題の各項目について行った(別表)．将来的に本格的な文献レビューへ発展させる際には，あらかじめ定めたプロトコルに従って網羅的検索を実施する．

C. 研究結果

C-1. 予備的に収集した文献の領域別整理

予備的に収集した 34 件を解析目的の領域別に整理すると，表 1 のようになった．ただし前述のとおり，これは網羅的検索による集計ではなく著者の選択に基づくため，領域ごとの多寡を定量的に論じることはできない．

表 1　予備的に収集した文献の領域別整理(計 34 件)

解析目的領域	件数	代表トピック
報告・評価基準	5	TRIPOD+AI 他
予測モデル方法論・報告品質	7	RoB，報告遵守，較正
因果推論	4	TMLE/AIPW/DML
NLP・大規模言語モデル	6	臨床文書，実装
サーベイランス・予測	7	早期警戒，流行予測
公平性・説明可能性	5	公平性ドリフト，XAI

C-2. 予備的に収集した文献から見た領域別の状況

(1) 予測モデル方法論・報告品質. この領域では，機械学習を用いた研究の報告・方法論的品質が低いことを，複数の系統的レビューが繰り返し示している．例えば Navarro らは 152 研究を評価し，その 87%が高いバイアスリスクにあると判定した．TRIPOD 項目の遵守率は中央値で約 39%，腫瘍領域に限ると約 41%にとどまる．われわれが特に問題と見るのは，識別能(discrimination)はほぼ全研究で報告される一方，キャリブレーションを報告する研究が 3 割前後しかない点である．Van Calster らはこれを「予測解析のアキレス腱」と呼んだ．外的検証を経ずに実装に至る例も少なくない．これらの欠落こそが，TRIPOD+AI(2024) や PROBAST+AI(2025)が整備された直接の背景をなしている．

(2) 因果推論. この領域でわれわれが中核と見たのは，交絡調整に用いるニュアンス関数(傾向スコアやアウトカム回帰)を機械学習で推定しつつ，妥当な統計的推論を保つ一群の二重頑健推定 — ターゲット最尤推定(TMLE)，拡張逆確率重み付け(AIPW)，二重/脱バイアス機械学習(DML) — である．これらは，機械学習の予測値を効果の推定式にそのまま代入したときに生じるプラグインバイアスを回避する．近年は，LLM を因果推論の副操縦士(co-pilot)として用いる構想も現れた．ただしわれわれは，因果的主張の妥当性が最終的に研究設計と識別仮定に依存するという原則は，AI の導入によって変わるものではないと考える．行政・介護データのような実務領域への応用報告は，現時点では乏しい．

(3) NLP・大規模言語モデル. LLM に関する研究は，総説や系統的レビューの形で急速に増えている．アーキテクチャの観点では，BERT 系が分類や情報抽出に，GPT 系が生成や対話に向くと整理される．応用はメンタルヘルス，臨床実装，医療者教育，グローバルヘルスへと広がった．一方でわれわれが注目したのは，実環境の医療ワークフローへの定着を実証した研究が依然として少なく，あるレビューでは該当する実証研究がわずか数件にとどまっていた点である．ハルシネーション，学習データに由来するバイアス，出力の非決定性といった LLM 固有のリスクへの対処と，評価方法の標

準化が課題として残る．

(4) サーベイランス・アウトブレイク予測. 感染症サーベイランスの領域では，AI の応用が具体的な成果として現れつつある．早期警戒システムを対象とする系統的レビュー(67 研究)が，手法・データソース・課題を整理している．BlueDot や EPIWATCH, HealthMap, GPHIN といった実例があり，COVID-19 の初期には，公的機関の確認に先立って異常な集積を検知した事例が報告された．米国 CDC も FluSight 等で機械学習を活用している．われわれが課題と見たのは，変異株の出現や政策変更といった構造変化に対して予測が脆弱になる点であり，これに対応する新たなモデリング手法や，LLM による多源シグナルの統合が試みられている．

(5) 公平性・説明可能性. この領域では，感受性属性で定義される部分集団の間で予測性能に差が生じることが問題となっている．公平性を測る統計的指標は複数あるが，それらは相互に両立しないことがあり，文脈に応じた価値判断を要する．近年は，配備後に公平性が経時的に劣化する「公平性ドリフト」が，モデル保守の新たな次元として論じられている．説明可能AI(XAI)は信頼の構築とバイアスの発見に資するが，われわれはそれが妥当性の検証に代わるものではないと考える．

C-3. 報告基準と方法論的品質の横断所見

報告・評価の枠組み (TRIPOD+AI, PROBAST+AI, TRIPOD-LLM, CONSORT-AI)は 2020 年以降に整備が進んだ．しかし複数の調査は，これらのガイドラインが公表された後も報告の遵守がさほど改善していないことを示している．識別能の報告は高頻度である一方，キャリブレーション，臨床的有用性，外的検証，そしてモデルを再現できる程度の記述は，依然として乏しい．これは AI を用いた予測モデル研究に共通する弱点として，領域を越えて観察された．

D. 考察

D-1. 主要知見

本予備的マッピングからわれわれが読み取ったのは，AI が医療・公衆衛生の統計解析の幅を広げる一方で，報告の透明性・方法論的厳密性・公平性の確保が領域を越えた共通の課題である，ということである．とりわけ予測モデリングでは，予測精度の追求に比して，キャリブレーション・外的検証・臨床的有用性の評価と報告が一貫して不足していた．

D-2. 研究ギャップ

以上を踏まえ，われわれは本領域に次のギャップがあると考える．

- 報告基準(TRIPOD+AI 等)の遵守は，ガイドライン公表後も十分に改善していない．
- キャリブレーション・臨床的有用性・外的検証の報告が，予測精度に比して乏しい．
- LLM の臨床実装を前向き・実環境で検証した研究が少なく，多くがレビュー段階にとどまる．
- 因果推論への機械学習の応用について，公衆衛生実務や行政データを対象とした報告が限られる．
- 配備後のモニタリング(公平性ドリフト・データセットシフト)を実証した研究が乏しい．
- 日本語文献および行政・介護データを対象とした研究のマッピングが，国際文献に比して乏しい．
- OCR と LLM を組み合わせた非構造化行政文書の構造化など，マルチモーダル統合の報告が限られる．

D-3. 限界

本研究にはいくつかの限界がある．第一に，本稿は，複数の学術データベース(PubMed, Embase 等)に対する再現可能な検索式の適用，重複除去，2 名独立スクリーニング，フロー図による特定から採択までの全数把握，という網羅的・体系的な文献レビュー(systematic review や scoping review)が要求する手順を，いずれも経ていない点である．

提示した 34 件および領域別の整理は予備的な把握であって，件数や分布を定量的な結果として解釈することはできない．第二に，対象言語を実質的に英語と日本語に限ったため，他言語の文献を取りこぼしている可能性がある．第三に，収集と整理が単一のレビューアによるため，選択の偏りを排除できていない．これらの限界はあるものの，本稿は本領域の概観と課題把握としては有用と考える．

E. 結論

われわれは本領域の文献を予備的に俯瞰し，報告の透明性，キャリブレーションと外的検証，公平性，そして実装の検証という，領域を越えて共通すると思われる課題を提示した．ただしこれらの所見は，予備的に収集した限定的な文献に基づくものであり，確定的な結論ではない．柔軟な計算手法がもたらす予測能力を，推定対象の明確な定義・仮定の開示・適切な検証・誠実な報告という統計的推論の枠組みのもとで運用すること ― これが本領域に求められる姿勢だとわれわれは考える．

F. 健康危険情報

特記すべき事項なし．

G. 研究発表

1. 論文発表：なし
2. 学会発表：なし

H. 知的財産権の出願・登録状況

1. 特許取得：なし
2. 実用新案登録：なし
3. その他：なし

引用文献

1) Tricco AC, et al. PRISMA-ScR. Ann Intern Med. 2018;169(7):467-473.
2) Peters MDJ, et al. Updated guidance for scoping reviews. JBI Evid Synth. 2020;18(10):2119-2126.
3) Collins GS, et al. TRIPOD+AI statement. BMJ. 2024;385:e078378.
4) Collins GS, et al. TRIPOD statement. BMJ. 2015;350:g7594.
5) Moons KGM, et al. PROBAST+AI. BMJ. 2025. (PMC11931409)
6) TRIPOD-LLM reporting guideline. Nat Med. 2025. doi:10.1038/s41591-024-03425-5.
7) Liu X, et al. CONSORT-AI extension. Nat Med. 2020;26:1364-1374.
8) Navarro CLA, et al. Risk of bias in ML-based prediction models: systematic review. BMJ. 2021;375:n2281.
9) Navarro CLA, et al. Completeness of reporting of ML-based prediction models. BMC Med Res Methodol. (PMC8759172)
10) Van Calster B, et al. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17:230.
11) Andaur Navarro CL, et al. Methodological conduct of ML-based prediction models. J Clin Epidemiol. 2022.
12) Dhiman P, et al. Reporting of ML-based prognostic models in oncology. (PMC8592577)
13) ML-based COVID-19 prognostic models: TRIPOD/TRIPOD+AI review. (PMC12866346)
14) Riley RD, et al. Minimum sample size for prediction models. Stat Med. 2019;38:1262-1296.
15) Moccia C, et al. ML in causal inference for epidemiology. Eur J Epidemiol. 2024;39:1097-1108.
16) Petersen M, et al. AI-based copilots to generate causal evidence. NEJM AI. 2024;1(12):AIp2400727.
17) Current Epidemiology Reports. AI tools for causal research. 2025;12:6.
18) La Mastra C, et al. AI to study causality in public health: systematic review. Eur J Public Health. 2024;34(S3).
19) LLM in mental health: scoping review. J Med Internet Res. 2025;27:e69284.
20) LLM architectures in health care: scoping review. J Med Internet Res. 2025;27:e70315.
21) Li HY, et al. Implementing LLMs in health care. J Med Internet Res. 2025;27:e71916.
22) LLMs in real-world clinical workflows: systematic review. Front Digit Health. 2025;7:1659134.
23) LLMs in global health. Nat Health. 2025/2026. doi:10.1038/s44360-025-00024-7.
24) Bakken S. Evaluation of LLMs in health and biomedicine. JAMIA. 2025;32(4):603.

25) Villanueva-Miranda I, et al. AI in early warning systems for infectious disease surveillance. Front Public Health. 2025;13:1609615.
26) AI in infection surveillance. (ScienceDirect 2025; S2319417025001039)
27) AI-driven epidemic intelligence. Front Artif Intell. 2025;8:1645467.
28) Borham A, et al. AI in epidemic watch. Front Digit Health. 2025;7:1692617.
29) CDC. Vision for Using AI in Public Health. 2025.
30) AI for modelling infectious disease epidemics. Nature. 2025;638(8051):623-635.
31) Quigley A, et al. EPIWATCH. Int J Infect Dis. 2025;152:107579.
32) Davis SE, et al. Emerging algorithmic bias: fairness drift. JAMIA. 2025;32(5):845-854.
33) Navigating fairness in clinical prediction models. (PMC12535043)
34) Naderalvojoud B, et al. Data biases, fairness and clinical utility. JAMIA Open. 2025;8(5):ooaf115.
35) Paulus JK, Kent DM. Predictably unequal. NPJ Digit Med. 2020;3:99.
36) Bias in AI for medical imaging. (PMC11880872, 2025)

別表　予備的に収集した文献の一覧(計 34 件)

注：本一覧は，一般のウェブ検索エンジンによる計 7 回の検索と著者の選択により予備的に収集した 34 件である．網羅的検索によるものではなく，件数や領域別分布を定量的に解釈すべきではない．著者名・誌名・識別子は本検索時に最終確認する．

No.	文献(第一著者 年 / 誌)	領域	AI 手法／トピック	主な知見・課題
1	Tricco 2018 (Ann Intern Med)	方法論	PRISMA-ScR	スコーピングレビュー報告基準
2	Peters 2020 (JBI Evid Synth)	方法論	JBI methodology	実施手順の更新指針
3	Collins 2024 (BMJ)	報告基準	TRIPOD+AI	回帰・ML 予測モデルの報告 GL, 2015 版を置換，公平性報告を組込
4	Collins 2015 (BMJ)	報告基準	TRIPOD	予測モデル報告 GL(22 項目)原版
5	Moons 2025 (BMJ)	報告基準	PROBAST+AI	開発 16・評価 18 設問の二部構成, ML 対応
6	TRIPOD-LLM 2025 (Nat Med)	報告基準	TRIPOD-LLM	LLM 研究の報告 GL
7	Liu 2020 (Nat Med)	報告基準	CONSORT-AI	AI 介入試験の報告拡張
8	Navarro 2021 (BMJ)	予測方法論	ML 予測モデル RoB SR	152 研究の 87%が高バイアスリスク, EPV 不足・欠測・過学習評価の不備
9	Navarro (BMC Med Res Methodol)	予測方法論	報告遵守 SR	TRIPOD 遵守 中央値 38.7%，モデル仕様・性能の報告稀
10	Van Calster 2019 (BMC Med)	予測方法論	キャリブレーション	キャリブレーションは予測解析のアキレス腱，評価・更新の必要
11	Andaur Navarro 2022 (J Clin Epidemiol)	予測方法論	方法論的 conduct SR	欠測・内的検証・キャリブレーション報告に課題
12	Dhiman (PMC8592577)	予測方法論	腫瘍 ML 報告 SR	TRIPOD 遵守 中央値 41%，報告の改善が急務
13	COVID prognostic SR 2025 (PMC12866346)	予測方法論	TRIPOD/TRIPOD+AI	ML 予後モデルの報告品質が回帰に劣後
14	Riley 2019 (Stat Med)	予測方法論	最小標本サイズ	開発に必要な EPV/標本サイズの算出法
15	Moccia 2024 (Eur J Epidemiol)	因果推論	TMLE/AIPW/DML	二重頑健推定の概説，プラグインバイアス回避
16	Petersen 2024 (NEJM AI)	因果推論	LLM co-pilot	因果エビデンス生成の AI 副操縦士
17	Curr Epidemiol Rep 2025	因果推論	AI 因果ロードマップ	因果推論各段階への AI 統合の機会
18	La Mastra 2024 (Eur J Public Health)	因果推論	公衆衛生因果 AI SR	複雑な交互作用のモデル化に期待
19	JMIR 2025 e69284	NLP/LLM	メンタルヘルス scoping	応用領域・性能・課題のマッピング
20	JMIR 2025 e70315	NLP/LLM	アーキテクチャ scoping	BERT=分類/抽出, GPT=生成/対話
21	Li 2025 (JMIR e71916)	NLP/LLM	臨床実装レビュー	モデル選択の指針，実装過程の支援
22	Front Digit Health 2025	NLP/LLM	実世界ワークフロー SR	実装の実証は限定的(4 研究)
23	Nat Health 2025/26	NLP/LLM	グローバルヘルス	低中所得国での生成 AI 応用

No.	文献(第一著者 年 / 誌)	領域	AI 手法／トピック	主な知見・課題
24	Bakken 2025 (JAMIA)	NLP/LLM	評価の前進	保健・生物医学での LLM 評価枠組み
25	Villanueva-Miranda 2025 (Front Public Health)	サーベイランス	早期警戒 SR	67 研究，手法・データ源・課題の整理
26	ScienceDirect 2025	サーベイランス	感染監視の AI	アウトブレイク予測・発生推定の改善
27	Front Artif Intell 2025	サーベイランス	疫学インテリジェンス	BlueDot 等，多言語 NLP による早期検知
28	Borham 2025 (Front Digit Health)	サーベイランス	epidemic watch	ML/DL による流行予測と資源配分
29	CDC 2025	サーベイランス	FluSight 等の ML	複数源統合による流行予測の改善
30	Nature 2025;638:623-635	サーベイランス	流行モデリングの AI	構造変化への予測脆弱性に対応する新手法
31	Quigley 2025 (IJID)	サーベイランス	EPIWATCH	公的確認に先立つ早期シグナル
32	Davis 2025 (JAMIA)	公平性/XAI	公平性ドリフト	配備後の公平性経時劣化，保守の新次元
33	PMC12535043	公平性/XAI	公平性指標の運用	指標の非両立, TRIPOD+AI で報告義務化
34	Naderalvojoud 2025 (JAMIA Open)	公平性/XAI	データバイアス	バイアスが臨床的有用性に及ぼす影響を評価