令和７年度　厚生労働行政推進調査事業費補助金　政策科学総合研究事業
分析ガイドラインの改定に向けた費用対効果評価における
方法論およびツール等の開発に関する研究

分担研究報告書

医療技術評価における生成 AI の信頼性・報告基準と今後の活用展望

森脇 健介
立命館大学 生命科学部 生命医科学科 医療政策・管理学研究室

近年、生成 AI・大規模言語モデル（LLM）は、システマティックレビューや経済モデル構築、RWD 解析など HEOR 業務の効率化に寄与している一方、透明性・再現性の不足、ハルシネーション、バイアスといった課題が指摘されている。本研究は、HTA 領域における AI 活用の信頼性確保に向け、国際的ガイドラインの動向を整理した。ISPOR は ELEVATE-GenAI フレームワークを提唱し、モデル特性や再現性、バイアス評価など 10 領域での報告を求めている。NICE は AI 活用をエビデンス生成、技術評価、業務効率化の 3 領域で推進し、方法論整備を進めている。CDA-AMC は AI を補助的手段と位置付け、透明性と人間の関与を重視している。これらを踏まえ、日本の HTA 制度では AI 利用の明示、リスクに応じた段階的運用、人材育成、国際整合的な方法論整備が必要である。AI は有用だが、透明性と検証可能性を担保した慎重な導入が求められる。

1. はじめに

近年、生成 AI および大規模言語モデル(LLM)は、医療経済・アウトカム研究(HEOR)において急速に活用され始めている。具体的には、システマティックレビュー(SLR)、医療経済モデル構築、リアルワールドデータ解析など、従来は人的労力を要していた作業の自動化・効率化が可能となっている。しかし、LLM の利用には以下の課題が存在する。すなわち、透明性の欠如、再現性の不足、ハルシネーション(誤情報生成)、バイアスや公平性の問題などである。HEOR・HTA 領域では LLM 利用の報告基準が未整備であり、研究の信頼性確保が課題となっている。本年度は、それらの「活用可能性」の段階を踏まえ、信頼性・透明性・再現性の確保を中心とした国際的ガイドラインの動向と、HEOR 実務への実装的展開を整理する。

2. 国際的な指針の整理

2-1. ISPOR の状況

ISPOR のワーキンググループは、HEOR 研究における LLM 利用を透明かつ再現可能に

報告するためのガイドライン ELEVATE-GenAI フレームワークを開発した[1]。このフレームワークは、LLM を用いた研究の報告標準、研究の再現性確保、HTA や政策判断での信頼性向上を目的としている。ELEVATE-GenAI フレームワークは、Bedi らによる AI 性能評価ドメイン、Holistic Evaluation of Language Models ベンチマークを基礎として構築されているが、ISPOR の Generative AI ワーキンググループは、HEOR 分野特有の方法論的および規制上の要件に対応するため、以下の 3 つの追加ドメインを導入した。

- モデル特性(Model characteristics)
- 再現性および一般化可能性(Reproducibility and generalizability)
- セキュリティおよびプライバシー(Security and privacy)

その結果、本ガイドラインは 10 の評価ドメインから構成される(表 1)。なお、各ドメインを以下の基準で評価し、すべてのドメインの点数を合計して総合スコアを算出することにより、各研究のガイドライン遵守度を半定量的に要約することもできる。

- 3 点：十分に報告されている(Clearly reported)
- 2 点：不明確または部分的(Ambiguous)
- 1 点：未報告(Not reported)

表 1. ELEVATE-GenAI チェックリスト

評価領域	チェック項目
モデル特性(Model characteristics)	モデル名、バージョン、開発者、公開日、ライセンス(オープンソースまたは商用)、モデルアーキテクチャが記述されているか。 学習データの情報(例：PubMed などのドメイン特化データ)およびファインチューニングの詳細が記載されているか。
精度評価(Accuracy assessment)	タスクに応じた精度指標(例：Precision、Recall、F1 スコアなど)が報告されているか。 出力結果が人間の評価やゴールドスタンダードデータセットと比較して検証されているか。
網羅性評価 (Comprehensiveness assessment)	出力結果が既存の研究レビューや検証済みモデルなどのベンチマークと比較され、網羅性が確認されているか。 重要な要素がすべて含まれていることについて専門家による評価が行われているか。

事実性検証(Factuality verification)	出力の事実性を確認する方法(例：文献との照合、専門家レビュー)が記述されているか。 不一致があった場合の修正方法や対応が記録されているか。
再現性および一般化可能性(Reproducibility protocols and generalizability)	再現性確保のための情報(例：トレーニングコード、プロンプト内容、ハイパーパラメータ)が公開されているか。 独立した研究者が検証できるワークフローが提示されているか。 同様の研究課題への一般化可能性について検討されているか。
ロバストネス検証(Robustness checks)	入力の誤字や曖昧な質問などへの耐性を評価するテストが実施されているか。 その条件下でのモデル性能の変化が報告されているか。
公平性・バイアス評価(Fairness and bias monitoring)	出力が性別、年齢、人種などに関するバイアスや偏見を含んでいないか評価されているか。 必要に応じて公平性指標(例：demographic parity)を用い、バイアスが確認された場合の対処が示されているか。
実装環境と効率性(Deployment context and efficiency metrics)	実装環境(ハードウェア、ソフトウェア、実行可能コードなど)が明確に記述されているか。 処理時間、スケーラビリティ、計算資源などの効率性指標が報告されているか。
キャリブレーションと不確実性(Calibration and uncertainty)	モデルのキャリブレーション方法(例：Expected Calibration Error)が記述されているか。 出力結果の手動確認が必要となる閾値(例：SLRでの曖昧ケース)が定義されているか。

セキュリティとプライバシー (Security and privacy measures)	セキュリティ対策(暗号化、匿名化、アクセス制御など)が説明されているか。 GDPR や HIPAA など関連法規への適合が示されているか(該当する場合)。 知的財産権や著作権への配慮が記載されているか。

ELEVATE-GenAI の各ドメインの内容や報告事項は表 2 に要約される。なお、各ドメインには、現時点における確立された指標または報告基準の現状を反映した成熟度レベルが割り当てられているが、これらは今後の状況変化に応じて、再検討・更新されるものである。ELEVATE-GenAI を公表されている SLR や CEA の研究事例に適用して、体系的に評価する試みも行われている[1]。

表 2. ELEVATE-GenAI のドメインの詳細

ドメイン	内容(概要)	報告すべき事項	指標成熟度
モデル特性 (Model characteristics)	モデルの基本属性(名称、バージョン、開発者、アクセス方法、ライセンス、アーキテクチャ、学習データなど)を記述する。	・モデル名、バージョン、開発者、公開日、ライセンス、アクセス方法を報告・モデル構造(例：Transformer)を説明・学習データやファインチューニングの内容を記載	高
精度評価 (Accuracy assessment)	モデル出力が正解や期待される結果とどれだけ一致しているかを評価する。	・人間の評価やゴールドスタンダードデータとの比較・適切な指標(Precision、Recall、F1、AUC など)の報告	中
網羅性評価 (Comprehensiveness assessment)	出力がタスクの全要素を網羅しているかを評価する。	・既存レビューやモデルなどのベンチマークとの比較・専門家による評価	高
事実性検証 (Factuality verification)	出力内容が事実に基づいているかを評価する。	・文献やデータソースとの照合・誤情報や不一致の修正方法の記録	高
再現性・一般化可能性 (Reproducibility and generalizability)	研究結果が再現可能か、他の研究課題に適用可能かを評価する。	・コード、プロンプト、ハイパーパラメータの公開・再現可能なワークフローの提供・他研究への適用可能性の説明	高
ロバストネス (Robustness checks)	入力の変化(誤字、曖昧な質問など)に対するモデルの耐性を評価する。	・誤字や曖昧入力などへのテスト・性能変化の報告	高
公平性・バイアス (Fairness and bias monitoring)	性別、年齢、人種などによる偏りがないか評価する。	・バイアス検出・公平性指標(demographic parity など)の使用	低

実装環境・効率性 (Deployment context and efficiency)	モデルの実装環境や計算効率を評価する。	・ハードウェア(GPU 等)とソフトウェア構成・処理時間、計算資源、スケーラビリティ	高
キャリブレーション・不確実性 (Calibration and uncertainty)	モデルが不確実性を適切に表現できているか評価する。	・キャリブレーション指標(Expected Calibration Error など)・手動レビューの閾値設定	低
セキュリティ・プライバシー (Security and privacy)	個人情報やデータ保護、知的財産を適切に管理しているか評価する。	・暗号化、匿名化、アクセス制御・GDPR/HIPAA 等への準拠・知的財産保護	低

ELEVATE-GenAI ガイドラインは、HEOR における LLM の利用状況を報告するための基本的な枠組みを提供するが、下記のような課題に留意した上で、今後、活用と議論を進めながら、継続的に見直しを行う必要がある。

- 枠組みの根拠となった対象を絞った文献レビューは体系的ではなく、関連する情報源が欠落している可能性がある。10 の領域は専門家の合意と文献統合によって導き出されたが、HEOR における LLM の利用状況に関するすべての関連側面を、不必要な複雑さや報告負担を招くことなく確実に捉えるためには、さらなる検証が必要である。
- 自己評価を支援するためにスコアリングシステムが試験的に導入されたが、その将来的な有用性は、より広範なユーザーからの意見に左右される。
- 特定のドメイン定義は概念的に類似しているため、一貫して適用することが難しい場合がある。
- 本フレームワークの HEOR タスク全体への一般化可能性については、更なる実証的検証が必要である。
- AI 研究で一般的に用いられる多くの評価指標は、経済評価モデルのパラメータ推定や健康状態の特定といった HEOR 特有のタスクでは検証されていない。
- 反復的または半自律的なタスクを実行するエージェント アプローチが普及するにつれて、この分野での追加のガイダンスが必要となる。

2-2. NICE の状況

NICE は、2024 年に「Statement of Intent for Artificial Intelligence(AI)」を公表し、医療分野および HTA における AI 活用の方針を示した[2]。同文書では、AI が医療におけるエビデンス生成、医療技術の評価、さらには HTA 機関自身の業務効率化において重要な役割を果たす可能性があることが指摘されている。

AI は機械学習などのデジタル技術を用いて、人間の知的活動を模倣するシステムであり、医療分野では臨床予測モデル、診断支援、デジタル治療などの形で利用が拡大している。NICE は、AI 技術が医療システムの課題解決に寄与する可能性を認識しており、特に COVID-19 後に顕在化した医療提供体制の負担、待機時間の増加、人材不足といった問題への対応において AI の活用が期待されるとしている。一方で、AI 技術の導入にあたっては、その有効性や費用対効果を含む価値を明確に評価する必要があり、HTA 機関としての役割が重要であるとされている。

NICE は AI に関する取り組みを以下の 3 つの優先領域に整理している。**第一は、AI を用いたエビデンス生成方法の整備**である。AI はシステマティックレビュー、エビデンス合成、経済評価などの研究プロセスにおいて、文献検索やデータ分析の効率化を可能にする。しかし、AI を用いた研究では、アルゴリズムの透明性、バイアスの管理、再現性、倫理的配慮などの課題が存在する。そのため NICE は、AI を用いたエビデンス生成に関するベストプラクティスのガイダンスを整備し、研究の信頼性を確保することを目指している。

第二は、AI を組み込んだ医療技術そのものの評価である。AI を用いた医療技術には、臨床予測モデル、画像診断 AI、デジタルヘルス技術などが含まれる。これらの技術の評価では、従来の医療機器評価に加え、アルゴリズムの妥当性、適応的アルゴリズムの更新、実臨床環境における性能、アルゴリズムバイアス、説明可能性、サイバーセキュリティといった新たな要素を考慮する必要があるとされている。NICE はこれらの課題に対応するため、AI 技術の評価基準や評価ツールの開発を進める方針である。

第三は、HTA 機関自身の業務における AI 活用である。NICE は、AI を用いることで文献レビューやデータ解析、ガイダンス作成などのプロセスを効率化できる可能性があると指摘している。ただし、AI の利用にあたっては透明性や説明責任、サイバーセキュリティ、リスク管理などを確保する必要がある。また、職員や委員会メンバーが AI 技術を適切に評価・活用できるよう、AI リテラシーの向上も重要とされている。

さらに NICE は、AI 評価の方法論の整備には国際協力が不可欠であるとして、HTA 機関や規制当局、研究機関との連携を強化する方針を示している。評価基準の国際的な調和を進めることで、AI 技術の評価における重複を減らし、効率的な方法論開発を進めることが期待されている。

以上のように、NICE は AI を医療技術評価における新たな重要課題として位置付け、AI を用いたエビデンス生成の方法論整備、AI 技術そのものの評価、HTA プロセスへの AI 活用という 3 つの観点から取り組みを進めている。

2-3. CDA-AMC の状況

Canada’s Drug Agency(CDA-AMC)は、2025 年に「Position Statement on the Use of Artificial Intelligence in the Generation and Reporting of Evidence」を公表し、HTA における AI の利用に関する基本方針を示した[3]。同声明は、医療技術評価に提出されるエビデンスの生成および報告の過程における AI 手法の利用について、透明性、厳密性、信頼性を確保する必要があることを強調している。

CDA-AMC は、近年 AI 技術が HTA 関連研究において広く検討されるようになっていることを踏まえ、AI の利用が適切かつ価値を付加する場合に限り、エビデンス生成を補助する手段として活用されるべきであるとの立場を示している。AI の利用により研究効率の向上が期待される一方で、アルゴリズムの透明性、倫理性、信頼性、適切性に関する懸念が存在するため、AI を用いた研究の評価には明確な指針が必要であるとされている。

同声明では、HTA における AI 利用の可能性として、主に以下の領域が示されている。**第一に、システマティックレビューおよびエビデンス統合の分野**である。AI は文献検索戦略の作成、研究デザインによる分類、文献スクリーニング、データ抽出などのプロセスを自動化する可能性がある。また、大規模言語モデルを用いてメタ解析に必要なコード生成やデータ合成を支援することも検討されている。

第二に、臨床エビデンスの生成である。AI は臨床試験の設計や患者選定、用量設定、試

験期間の最適化などに活用される可能性がある。また、電子カルテなどのデータを自然言語処理によって解析し、試験対象患者の特定や副作用の検出に利用することも想定されている。さらに、機械学習を用いることで複雑な共変量関係を考慮した統計解析や、合成対照群の作成などが可能になるとされている。

第三に、リアルワールドデータ(RWD)の解析である。AI は電子カルテ、レジストリ、行政データベースなどの大規模データを統合・標準化する過程において、データクレンジング、重複除去、欠損値補完などを自動化する役割を果たす可能性がある。また、機械学習による特徴量選択や因果推論手法の活用により、治療効果の推定精度を向上させることが期待されている。

第四に、費用対効果評価における経済モデルの開発である。AI は疾患進行や医療費構造などの複雑なデータ解析を通じて、モデル構造の設計やパラメータ推定を支援する可能性がある。また、大規模言語モデルを用いて経済モデルのコード生成やモデル更新を自動化する可能性も指摘されている。これにより、複数のモデルを比較することで構造的不確実性を検討することが可能となる。

一方で CDA-AMC は、AI の利用に伴うリスクにも注意を払う必要があるとしている。具体的には、アルゴリズムバイアス、透明性の低下、サイバーセキュリティリスク、人間の監督の低下などが挙げられる。そのため、AI は人間の判断を代替するものではなく、あくまで意思決定を補助する手段として利用されるべきであり、いわゆる「human-in-the-loop」の原則を維持する必要があるとされている。また、AI を利用した場合にはその使用方法を明確に開示し、モデルの検証結果やリスク管理方法を報告することが求められる。

さらに、AI の利用にあたっては、倫理的原則として人間の安全と福祉の促進、公平性の確保、責任の明確化、持続可能性の確保などを重視する必要があるとされている。CDA-AMC は今後、AI を用いたエビデンス生成の実例を継続的にモニタリングし、HTA の方法論ガイドラインの更新に反映していく方針を示している。

以上のように、CDA-AMC の声明は、AI を HTA におけるエビデンス生成の補助的手段として位置付けるとともに、その利用に際して透明性と厳密性を確保することの重要性を示している。

3. 日本の費用対効果評価制度への示唆

本研究では、HTA における AI 活用に関する国際的な指針の策定状況を整理した。AI の利用が今後拡大することを踏まえ、HTA における AI 活用の方法論や評価基準の整備が国際的に進展していくことが期待される。本検討の結果を踏まえ、日本の HTA 制度において今後検討すべき事項として、以下の点を提言する。

(1) HTA 提出資料における AI 利用の透明性確保

費用対効果評価等の HTA 提出資料において、AI の利用有無および利用範囲を明示的に申告する仕組みを整備する。特に以下の工程における AI 利用について記載を求めることが望ましい。

- 文献検索・文献スクリーニング
- データ抽出および整理
- RWD/RWE 解析
- 統計解析および因果推論
- 経済モデル構築・コード生成
- 報告書作成

また、AI ツール名、バージョン、使用方法、人間による確認手順、検証方法、既知の限界などを記載する様式の整備を検討する。

(2) AI 利用に関するリスクに応じた段階的運用

HTA における AI 利用については、用途ごとのリスクに応じた段階的運用を行う。例えば以下のような区分が考えられる。

- **低リスク用途（導入しやすい）**
  - 文献検索補助
  - 文献分類
  - 図表整形
  - 要約作成
- **高リスク用途（厳格な検証が必要）**
  - 治療効果推定
  - 因果推論
  - RWD 分析
  - 経済モデル構造の設計
  - 費用対効果結果の算出

なお、高リスク用途では、独立再現、感度分析、代替手法との比較、人間による最終確認を必須とすることを検討する。

(3)　公的分析実施機関における AI 活用の検討

公的分析実施機関自身の業務においても、AI を活用した以下の効率化を検討する。

- 文献レビューの効率化
- データ整理・分析の支援
- HTA レポート作成の補助

ただし、AI 利用にあたっては透明性、説明責任、著作権、サイバーセキュリティなどのリスク管理を徹底する必要がある。

(4)　AI 評価能力を有する人材育成

製造販売業者だけでなく、公的分析実施機関、費用対効果評価専門組織、事務局においても AI を適切に評価できる人材の育成が必要である。特に以下の分野に関する能力強化が求められる。

- AI アルゴリズムの基礎理解
- 再現性確認
- データガバナンス
- AI 倫理
- サイバーセキュリティ

(5)　国際動向を踏まえた方法論整備

NICE、CDA-AMC 等の HTA 機関では AI 利用に関する方法論整備が進められている。日本の HTA 制度においても、国際的な方法論やガイドラインとの整合性を確保しながら制度整備を進めることが重要である。

4. さいごに

AI は HTA におけるエビデンス生成および評価プロセスの効率化・高度化に寄与する可能性を有する。一方で、透明性、再現性、バイアス、セキュリティなど新たな課題も生じるため、日本の費用対効果評価制度においては AI を全面的に導入するのではなく、透明性と検証可能性を確保した上で段階的に活用を進めることが望ましい。

5. 参考文献

1. Fleurence, R.L., et al., ELEVATE-GenAI: Reporting Guidelines for the Use of Large Language Models in Health Economics and Outcomes Research: An ISPOR Working Group Report. Value Health, 2025. **28**(11): p. 1611-1625.
2. NICE statement of intent for artificial intelligence (AI) 2024; Available from: www.nice.org.uk/corporate/ecd12

3. Canada’s Drug Agency Position Statement on the Use of Artificial Intelligence in the Generation and Reporting of Evidence. 2025; Available from: https://www.nice.org.uk/about/what-we-do/our-research-work/useof-ai-in-evidence-generation--nice-position-statement.