
AI 시대의 HTA:

Systematic review에 AI가 어떻게 활용되고 있나?

2025.11.28.
발표자: 설아람



This study was supported by the National Evidence-based Healthcare Collaborating Agency (grant no. NECA-A-24-014).

목차

1. 들어가며

AI 기술의 발전과 새로운 딜레마

- ▶ AI 기술 개발 속도는 이전 어떤 기술보다 빠르게 확산 중
- ▶ 대규모 언어모델(**large language model, LLM**)의 등장으로 고비용 학습데이터 없이도 텍스트를 '이해'하고 처리하는 시대 도래
- ▶ 그러나, 다수의 AI 도구(기계학습·딥러닝·NLP·생성형 AI 등)는 **투명성 부족, 가정과 한계의 불명확성, 일반화 문제**를 지님
- ▶ 객관성·신뢰성·타당성을 검증할 근거기반도 미흡
- ▶ AI의 발전은 빠르지만, 그 신뢰를 담보할 과학적·윤리적 기반은 뒤따르지 못하고 있음

[출처: Thomas 2025]

생성형 AI의 등장과 HTA의 변화

- ▶ LLM을 포함한 생성형 AI는 HTA의 세 축에 근본적인 변화를 가져올 잠재력이 있음
- **문헌 근거 합성(SR 자동화)**: 검색 전략 설계, 초록 선별, 데이터 추출, 메타분석 코드 작성 등 → 근거 합성 과정의 효율성과 재현성 향상
- **실사용근거(real-world evidence, RWE) 생성**: 임상기록·의료영상 등 비정형 데이터의 대규모 분석 가속화 → 실제 진료 환경의 근거 창출 확대
- **보건 경제 모델링**: 모델 개념화, 매개변수 추정, 검증 과정의 자동화 지원 → 신속하고 반복 가능한 경제성 평가 가능

[출처: Fleurence 2025]

생성형 AI의 한계와 인간의 감독 필요성

▶ 주요 한계 영역

- **과학적 타당성과 신뢰성:** 결과의 재현성과 검증이 충분하지 않음
- **편향 및 형평성 문제:** 학습데이터의 불균형이 평가 결과에 영향을 미칠 수 있음
- **규제 및 윤리적 고려:** 법적 책임·데이터 보호 등 제도적 틀이 미비

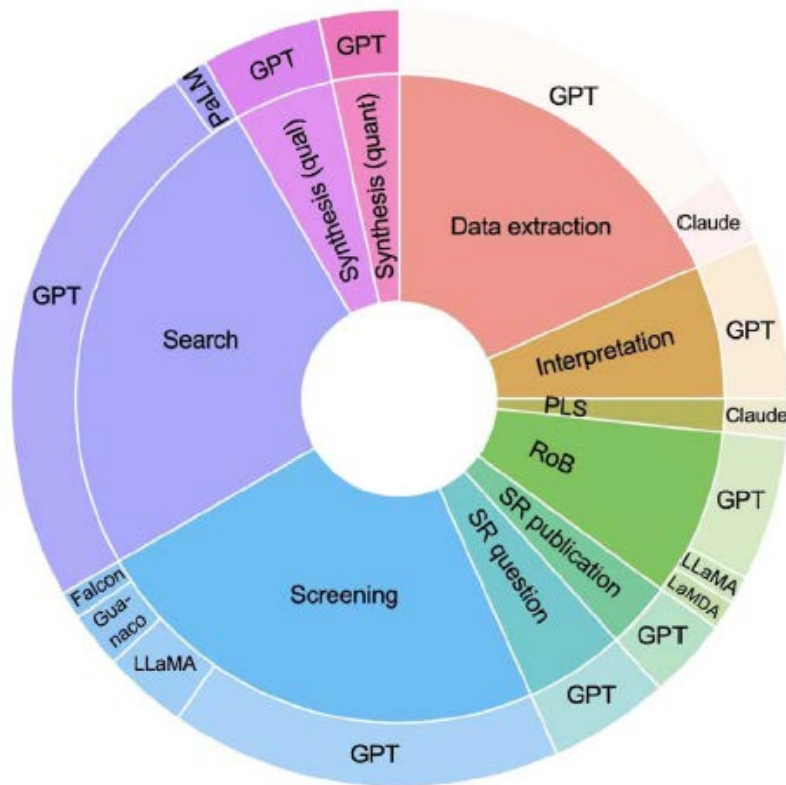
▶ 생성형 AI는 여전히 **초기 단계**에 있으며, **신뢰성·공정성·윤리성** 측면에서 다양한 한계를 지님 → HTA 적용 시 **지속적인 검증과 인간의 감독**이 필수적임

[출처: Fleurence 2025]

2. AI의 체계적 문헌고찰(SR) 단계별 활용

SR 단계별 AI 활용 현황 (파이-도넛 차트)

AI가 SR 각 단계 중 어디에 가장 집중 적용되고 있는가?



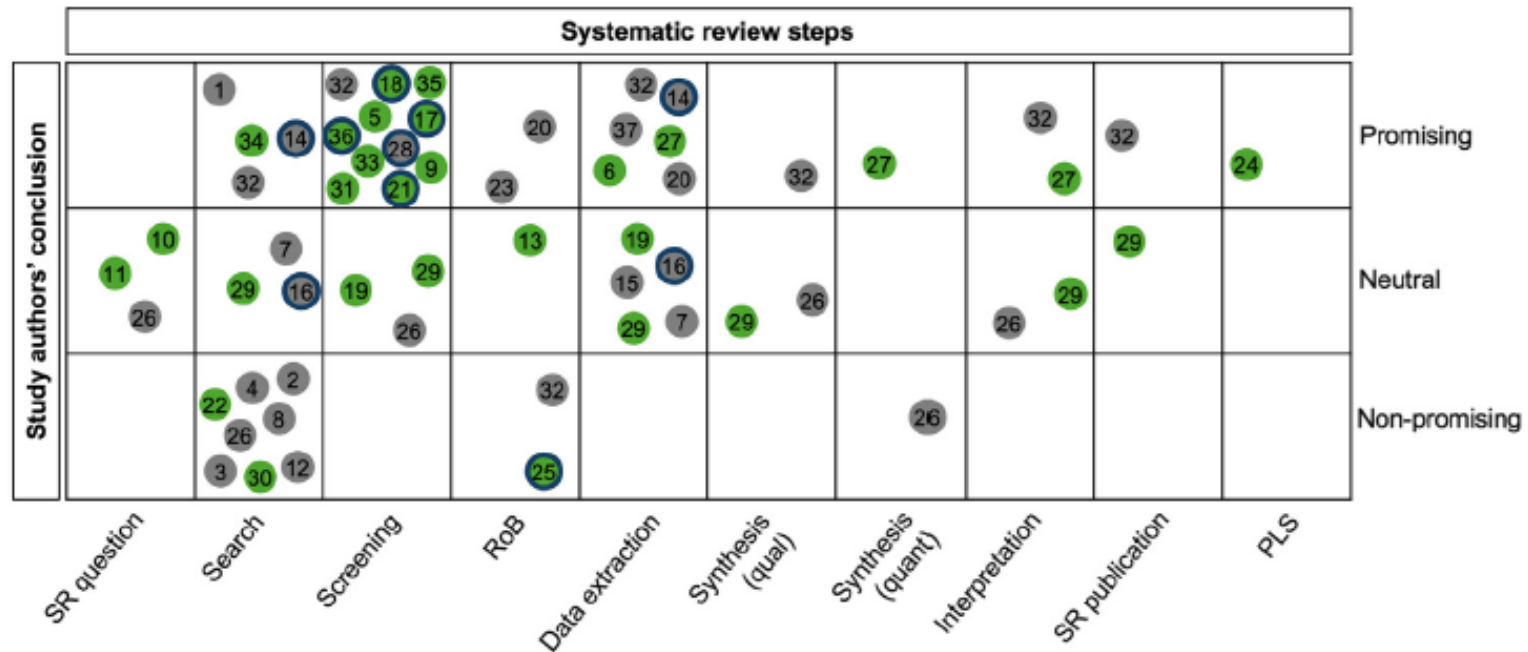
포함 연구 37개에서 추출된
전체 개별 SR 단계(n = 60)
▶ 중앙(내부 원형) : 전체 개별
SR 단계의 빈도와 비율
▶ 외곽 도넛: 각 단계에 사용된
LLM 유형의 백분율 분포

- 검색: 25%
- 선별: 23.3%
- 데이터 추출: 18.3%
- 비둘림 위험 평가(RoB): 8.33%
- 해석: 6.7%
- 질적 합성: 5%
- 양적 합성, SR 출판: 3.3%
- 일반어 요약(PLS): 1.7%

[출처: Lieberum 2025]

연구 설계별 SR 단계-결과 분포 (버블 차트)

AI 연구의 설계와 단계별 결론의 다양성: 유망함 / 중립적 / 비유망



- ▶ 주요 연구 설계와 각 SR 단계(x축)에 대한 연구자 결론 범주(y축)
- 녹색: 유효성 검증 연구(validation) / 회색: 기타 연구 설계 / 파란색 원: 사전 출판 논문(preprint)

[출처: Lieberum 2025]

체계적 문헌 검색 단계

- ▶ LLM은 효과적인 **검색 전략의 개발 및 최적화**를 지원하고, 회색 문헌 탐색에도 기여할 수 있음
- ▶ 자연어 처리 기술을 통해 대규모 의학 문헌을 빠르게 처리하여 **검색 효율성 향상**에 기여
- ▶ LLM은 초기 탐색에는 유용하지만, 엄격한 **SR 검색을 완전히 대체하기는 어렵다**고 제시
(Cochrane(2025.1))
- ▶ LLM이 설계한 검색 전략은 **민감도-정밀도 균형** 조정이 어렵고, 오류 검출을 지원하더라도 **전문가 검토를 대체할 수 없음**
- ▶ 결론: **AI는 출발점, 정보전문가의 검증이 품질 담보**

[출처: Sul 2025]

체계적 문헌 검색 단계

- ▶ CORE Information Retrieval Forum 2025 주요 결과
 - 참여자: 131명 정보전문가 (10개국, 주로 영국)
 - AI 논의 비중: 6개 주제 중 AI가 가장 활발
 - 긍정적 인식: 검색어 확장, 시각화, 요약 등 **반복 업무 효율화**
 - 주요 우려: 신뢰성, 재현성, 편향, '블랙박스' 문제, 비판적 사고 약화
 - 필요사항: 윤리적 가이드라인, 실무 교육, 표준화된 적용 모델

[출처: Eastaugh 2025]

체계적 문헌 검색 단계 – 최신 근거(1)

▶ 연구 개요

- LLM 기반 도구(Elicit AI)가 문헌 검색을 얼마나 잘 수행하는지를 실제 4개의 SR 사례를 통해 평가

▶ 주요 결과

- 평균 민감도(sensitivity): 39.5% (25.5-69.2%) → 기존 검색(94.5%)에 비해 낮음
- 평균 정밀도(precision): 41.8% → 기존 검색(7.6%)에 비해 낮음
- 기존 검색으로 누락된 일부 연구를 추가로 발견함
- 검색 재현성(reproducibility) 및 투명성 부족
- 결론: **보조 도구**로 유망, 주 검색 방법으로는 부적절

[출처: Lau 2025]

체계적 문헌 검색 단계 - 최신 근거(2)

▶ 연구 개요

- AutoLit Smart Search AI는 연구 질문을 입력하면 PICO 기반 Boolean 검색식을 자동 생성함

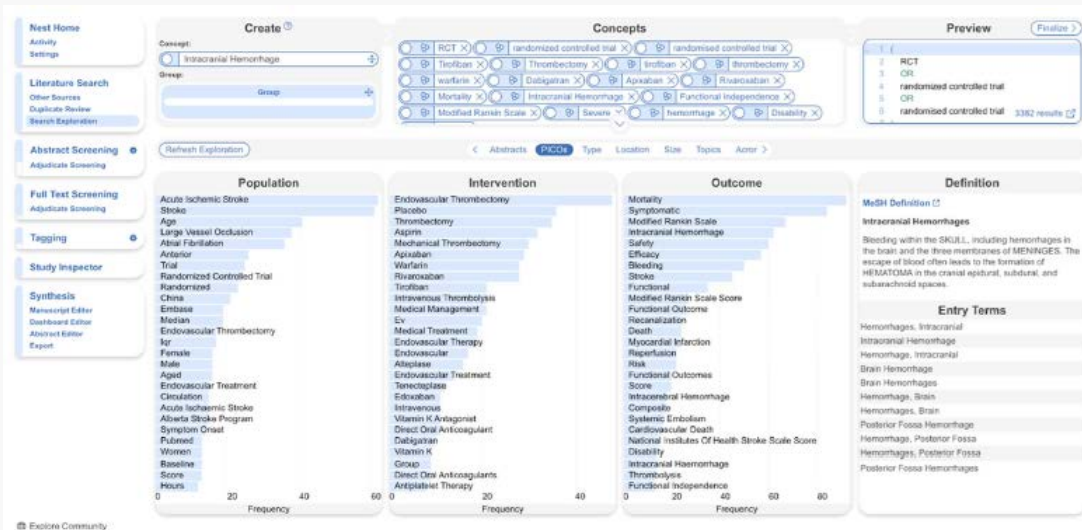


FIGURE 1 | Search Exploration predicts and maps Populations, Interventions, Comparators, and Outcomes from underlying abstracts resulting from Boolean PubMed queries, including reporting frequency and enabling adoption of new search terms.

[출처: Kallmes 2025]

체계적 문헌 검색 단계 – 최신 근거(2)

▶ 주요 결과

- Cochrane 검증세트(10개 리뷰) 기준 **검색 재현율 (recall) 76.8%**, 추가 검증세트(19개 리뷰)에서 **79.6%** 달성
- 이는 GPT-4.0 단독 수행(recall 13%)보다 현저히 우수하며, 인간 전문가 수준(70~85%)에 근접
- 정확도(Precision)는 1-5% 수준으로 전문가 검색과 유사, Cochrane 다중 DB 검색 대비 PubMed 단독 기반에서도 일관된 성능

[출처: Kallmes 2025]

문헌 선별 단계

- ▶ LLM은 대규모 문헌(제목·초록·키워드)의 **신속한 스크리닝**에 활용 가능
- ▶ **중복 문헌 제거** 및 필요시 **전문(Full-text) 스크리닝 자동화** 지원 잠재력 있음
- ▶ ML·AI 기반 도구는 효율성을 높이지만, 실제 SR에서는 **도입 속도가 더딤**
- ▶ **유효성 검증된 중단 규칙(stopping rules)** 부재가 주요 한계
- ▶ LLM은 **탐색적 수준**으로, 전문가 감독 아래 **보조적 도구로 권장**(Cochrane(2025.2))

[출처: Sul 2025]

문헌 선별 단계 – 최신 근거(1)

▶ 연구 개요

- GPT-4를 실제 정책 분야 체계적 문헌검토에 활용한 첫 실증 연구
- 제목·초록 스크리닝 단계에서 LLM 실제 성능 평가

▶ 주요 결과

- GPT-4가 전체 43,480편 중 **96%를 자동 제외**, 거짓 제외율 0.0-0.05% (추정 < 1%)
- **포함 판단 오류율 65%** → 포함 결정(inclusion)에는 부적합
- **작업 시간 약 30 근무일 단축** 효과 추정

[출처: Rubinstein 2025]

문헌 선별 단계 – 최신 근거(2)

- ▶ **GREP-Agent (Generative-AI-assisted Review Pipeline)**
 - 여러 LLM(예: GPT4o, Llama 3.1, Phi 4)을 서로 다른 역할의 에이전트로 구성해 문헌 선별 단계에서 자동화와 인간 검증을 병행하도록 설계 (여러 LLM 에이전트가 서로 검증하며 불확실한 문헌만 인간에게 넘김)

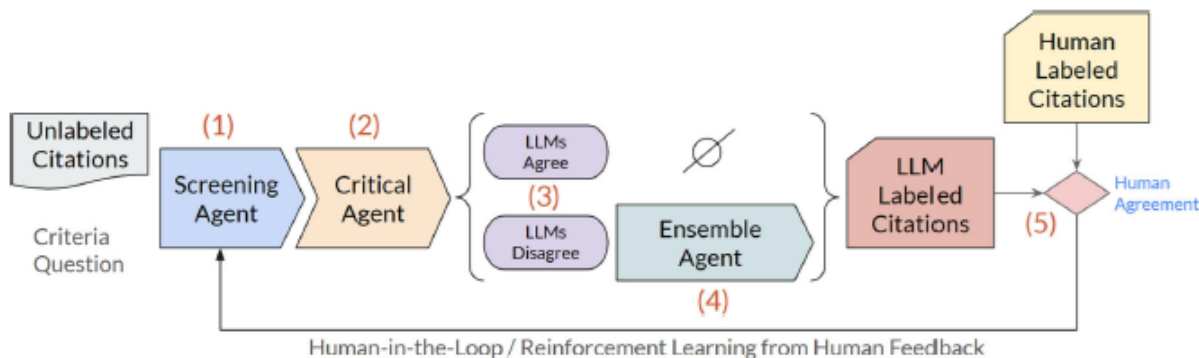


FIGURE 2 | The GREP-Agent architecture. Three LLM agents make up the core of our architecture; [1] the screening agent, [2] the critical agent, and [3] the ensemble agent.

[출처: Hu 2025]

문헌 선별 단계 – 최신 근거(2)

▶ 주요 결과

- 평가 데이터: 홍역(measles) SR 2,000편 이중 선별
- 민감도(recall) 86.4~95.3% / 작업량 71~92% 감소
- 프롬프트 문장만 바뀌도 성능 $\pm 6\%$ 차이 → **프롬프트 설계 중요**

▶ 결론: AI는 감독 가능한 자동 선별 시스템으로 발전하고 있음

- 단일 LLM → '협업형 LLM + 인간 보정 시스템'
- AI 성능 향상: 인간 검증가능한 품질(auditable performance) 수준

[출처: Hu 2025]

데이터 추출 단계

- ▶ AI, 특히 LLM은 문헌에서 필요한 정보를 **구조화·자동 추출**할 수 있음
- ▶ 연구의 기본 정보, 환자 특성, 주요 의료결과 (outcome) 등 **다양한 데이터 유형** 처리 가능
- ▶ LLM을 이용할 경우 데이터 추출 시 **정확성과 효율성 향상**이 가능함(Cochrane(2025.3))
- ▶ **프롬프트 설계(prompt engineering)** 전문성이 확보되면 초기 추출단계에서 **인간의 일부 역할을 보완**할 가능성 있음
- ▶ **결과 검증(human oversight)**은 여전히 필수

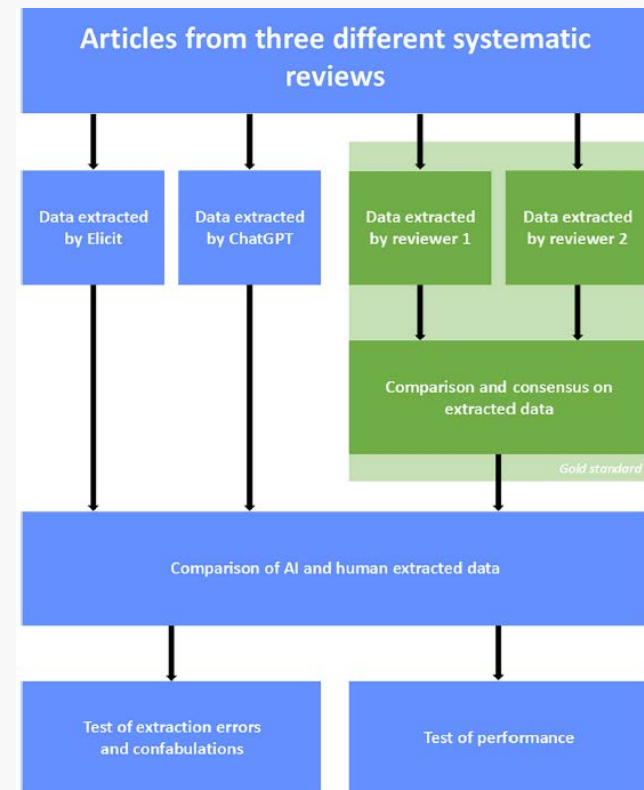
[출처: Sul 2025]

데이터 추출 단계 - 최신 근거(1)

[AI의 데이터 추출 성능 비교]

▶ 연구 개요

- Cochrane SR 3건(총 30편)을 대상으로 **ChatGPT-4o vs Elicit vs 인간 이중추출자** 비교
- 추출 항목: 연구 설계, 모집단 특성, 주요 변수
(Reviewer-specific variable)



[출처: Andersen 2025]

데이터 추출 단계 - 최신 근거(1)

▶ 주요 결과

- **정확도(F1)** : Elicit 92%, ChatGPT 90% - 두 도구 모두 인간 대비 **동등 수준**
- **오류(confabulation)** 4%(Elicit), 3%(ChatGPT), **주요 오류 유형**은 표/연속문 내 데이터 누락 또는 잘못된 라벨링
- 통계적으로 유의한 성능 차이 없음($p > 0.4$)

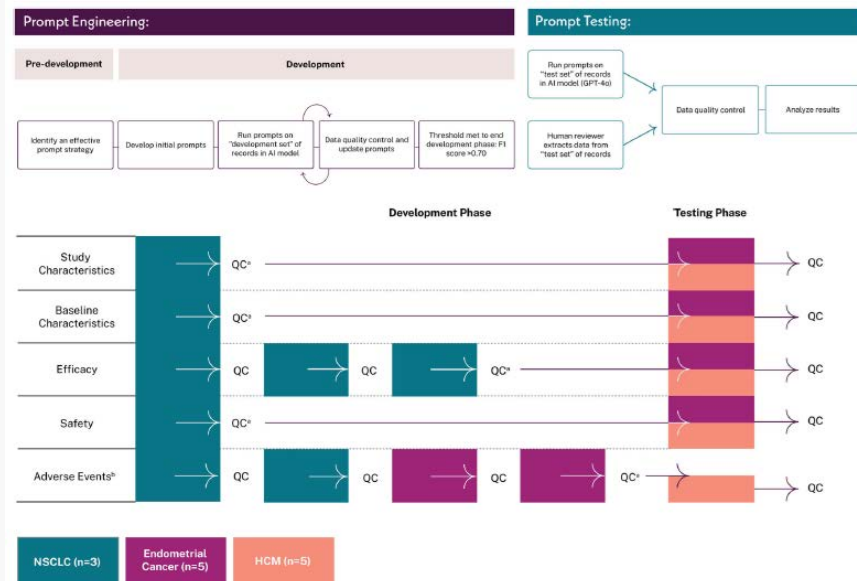
[출처: Andersen 2025]

데이터 추출 단계 - 최신 근거(2)

[프롬프트 설계가 좌우하는 LLM 추출 효율]

▶ 연구 개요

- **GPT-4o**를 이용해 3개 질환군(NSCLC, 자궁내막암, 비대심근증) RCT 15편에서 **프롬프트 엔지니어링 효과 분석**
- **데이터 규모: 6만개 이상 데이터 필드** (연구특성, 기저특성, 효능, 이상반응)



[출처: Murton 2025]

데이터 추출 단계 - 최신 근거(2)

▶ 주요 결과

- 단순 구조 항목(연구/기저 특성) $F1 > 0.85$, 복잡 항목(효능·안전성)은 $0.22 \sim 0.50$
- 프롬프트를 **질환 간 전이(testing on unseen data)** 시, 일부 영역에서 인간과 동등하거나 우수

▶ 결론 및 시사점

- **프롬프트 구조**가 성능 핵심 요인: 단일 복합 프롬프트는 효율적, 세부 병렬형은 정확도 향상·시간 증가
- 복잡 데이터에서 **혼동 및 과잉추출** 위험 지속
- 인간 검증(human-in-the-loop)과 질환별 미세조정(prompt refinement)이 성능 개선의 관건

[출처: Murton 2025]

데이터 추출 단계 - 최신 근거(3)

▶ 연구 개요

- 목적: 공중보건 근거검토에서 LLM 을 활용해 다양한 연구유형(실험·관찰·질적 연구 등)의 데이터 추출 가능성 평가
- 모델/환경: **RAG 기반 LLaMA 3-70B 모델** (UKHSA 보안 인프라 내 운영)
- 비교대상: **COVID-19 비약물적 중재 효과 리뷰**(173 필드, 24편 논문) → 인간 표준 대비 LLM 출력 정확성 비교

[출처: Simmons 2025]

데이터 추출 단계 - 최신 근거(3)

▶ 주요 결과

- **정확도(accuracy)**: 173개 필드 중 42 % 정확, 28 % 경미 오류, 21 % 주요 오류 → **수용가능성(acceptability)** 68 %
- **신뢰도(reliability)**: 10회 반복 시 평균 일치율 0.71 (“연구설계” 0.94 > “결과” 0.55)
- 오류특성도: 문자열 인식 한계, 기간 필드 오해, 출력 과다; Hallucination 1.2 % 로 낮음
- 리뷰어 의견: LLM 단독 보다는 **1차 자동 추출 + 인간 2차 검증(human-in-the-loop)** 방식이 현실적

[출처: Simmons 2025]

비둘림 위험의 평가 단계

- ▶ LLM은 가이드라인 문서를 기반으로 **RoB 평가 관련 정보**를 신속히 추출 가능
- ▶ Cochrane *RoB 2* 및 *ROBINS-I* 기준 적용 시 **부분 자동화된 평가 지원**이 가능함
- ▶ 하지만 **연구 간 정확도 편차**가 크고, **일관성 부족**이 문제로 지적됨
- ▶ “**현재 LLM 기반 RoB 평가는 권장되지 않음**”으로 결론(Cochrane(2025.4))
- ▶ 향후 특정 영역 중심의 **보조적 활용**만 가능할 것으로 전망

[출처: Sul 2025]

비둘림 위험의 평가 단계 - 최신 근거(1)

[LLM vs 인간 RoB 평가 비교: Cochrane 분석 결과]

▶ 연구 개요 및 주요 결과

- **ChatGPT-4o vs Cochrane 저자 비교: RoB 2 도구 적용 RCT 84편**
- **전반 일치 $\kappa=0.51$ (보통) / 민감도 53%, 특이도 99%**
- 도메인별 일치 범위 fair~moderate (선택 결과 κ 0.20 ↔ 측정 결과 κ 0.59)
- 평가 시간 대폭 단축(약 15% 이내)

▶ 결론 및 시사점

- **일관성·재현성 문제 및 데이터 의존성 여전**
- **설명가능성 확보 전까지는 보조적 도구로 한정**
- **향후 표준 프롬프트 및 도메인별 모델 튜닝 필요**

[출처: Taneri 2025]

비둘림 위험의 평가 단계 - 최신 근거(2)

[LLM의 RoB 평가 성능: 무작위배정 연구 근거]

- ▶ 연구 개요 및 주요 결과
 - ChatGPT-4o를 활용해 **Cochrane RCT 100편 중 75편**의 RoB 평가를 인간 합의판정과 비교
 - 전반적 일치율 **50.7% ($\kappa=0.22$)**, 무작위배정 영역은 **78.7%($\kappa=0.54$)**로 “fair~moderate” 수준
 - ChatGPT-ChatGPT 자체 일치 **74.7%**, 인간-인간 합의(16-81%)와 유사 수준
 - 단순 프롬프트가 복잡한 질문 방식보다 성능 우수, 판정시간 수초 이내
- ▶ 결론: RoB 도메인 중 '무작위배정'처럼 객관적 항목은 자동화 가능성 있음 (**보조적 활용은 가능**)

[출처: Rose 2025]

비둘림 위험의 평가 단계 - 최신 근거(3)

▶ 연구 개요

- **Claude 2** LLM을 활용한 **RCT의 RoB 평가 성능**을 정량적으로 검증한 최초의 대규모 실증연구

▶ 주요 결과

- 평균 일치율: 41-71% (D3: '결측자료' 가장 높음)
- Cohen's κ : 0.10 (D5 선택적 보고) ~ 0.31 (D3 결측자료)
- 주요 불일치 원인: 무작위배정 은폐·눈가림 해석 차이, 프로토콜 탐지 실패
- 민감도 분석: prompt 유형·모델 변경 시에도 fair 수준 이하

[출처: Eisele-Metzger 2025]

근거 합성 단계

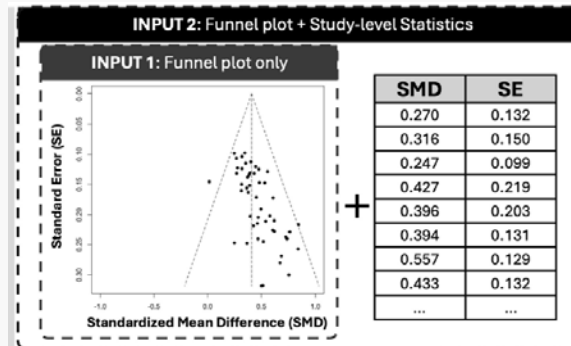
- ▶ AI는 여러 연구 데이터를 통합·분석하는 **근거 합성 단계**의 효율성 향상에 기여
- ▶ 복잡한 **네트워크 메타분석** 수행 시 자동화 및 계산 속도 향상 가능
- ▶ 대규모 데이터 분석을 통해 **패턴·추세·잠재적 교란 요인(confounding)** 탐지 가능
- ▶ 그러나 **결과의 신뢰성과 투명성 확보**는 여전히 도전 과제
- ▶ 결론: AI는 **속도·규모·범위**를 확장하지만, **최종 판단은 인간의 검증**을 거쳐야 함

[출처: Sul 2025]

근거 합성 단계 – 최신 근거(1)

▶ 연구 개요

- **GPT-4o와 Llama 3.2 Vision**을 이용해 **출판편향 (publication bias, PB)** 자동 탐지 시도
- 퍼널플롯 단독 vs. 퍼널플롯 + 통계입력(효과크기, 표준오차) 비교
- PB 유무, 표본 크기, 이질성 조건을 다양하게 설정한 시뮬레이션 연구



GPT 4o

Llama

Llama 3.2 vision

Prompt Engineering

Language-based prompt inspired by the
Cochrane Handbook

[출처: Xing 2025]

근거 합성 단계 – 최신 근거(1)

▶ 주요 결과

- **GPT-4o**: PB 없음 상황에서 **특이도 높음**, 대규모 연구(>50편)에서 안정적
- **Llama 3.2**: PB존재 시 **민감도 높음** (소표본에서 탐지력 우수)
- 전통적 검정(Egger 등) 대비, **소표본(15편)**에서 파워 0.50 vs. 0.15로 상대적 우위

▶ 결론: AI는 메타분석 단계에서 **패턴 인식과 속도**를 개선하지만, **정확한 해석은 전문가 검증이 필수**

[출처: Xing 2025]

3. 책임 있는 AI 활용과 도전 과제

AI 사용 지침(RAISE)

- ▶ Responsible AI in Evidence Synthesis
 - 여러 관련 기관 및 International Collaboration for Automation in Systematic Reviews (ICASR) 운영 그룹 회원들의 협력을 통해 개발
 - **공동체 기반 best practices** 구축을 목표로 함
 - **연구윤리 원칙**에 기반하여 근거 합성 생태계 전반에서 **AI의 책임 있고 투명하며 윤리적인 활용**을 보장
 - 연구자-기관-조직이 상호 협력하는 커뮤니티를 통해 **AI로 지원되는 신뢰할 수 있는 근거 합성 환경 조성** 추구

[출처: Thomas 2025]

AI 사용 지침(RAISE)

구분	주요 내용	핵심 초점
RAISE 1	근거 합성 생태계 내 8개 주요 역할별 권고안 제시 : 근거합성자, 방법론자, AI 개발자, 출판사, 연구비 지원기관 등	역할별 책임과 윤리 원칙 명확화
RAISE 2	AI 근거합성 도구의 구축·평가 가이드라인	도구 개발의 투명성·재현성 확보
RAISE 3	AI 도구의 구축·평가 가이드라인	안전한 활용과 윤리적 판단 기준 제시

[출처: Thomas 2025]

DESTINY 프로젝트

- 책임 있는 AI 기반 Living Evidence 혁신

[**D**igital **E**vidence **S**ynthesis **T**ool **I**Nnovation for
Yielding Improvements in Climate & Health]

- ▶ AI의 책임 있는 활용 모델로 주목받는 프로젝트
 - 영국 Wellcome 재단 지원, 국제 다기관 컨소시엄(9개 핵심 기관, 30여 협력기관, 총 45명 이상 연구자)
- ▶ 비전(Vision)
 - AI를 활용해 **기후·건강 분야의 근거생성 속도·품질·적시성을 혁신**
 - “사람 중심의 기술 공동개발(co-production)”을 핵심 가치로 설정

(출처: DESTINY 홈페이지)

환경 지속가능성에 대한 고려(1)

▶ 생성형 AI와 데이터센터: '전기 먹는 하마'

- 챗GPT 검색: 일반 웹 검색 대비 약 5배 전력 소모
- 탄소 배출 증가: 구글 ('19→'24, +51%), 마이크로소프트 ('20→'24, +23%)
- 국제에너지기구(IEA) : 향후 5년 내 AI 데이터센터 전력 수요 2배 이상 증가 전망
- 데이터센터 전력·수자원 사용 급증 → 지역 에너지 비용 상승, 지하수 고갈 등 환경 피해 우려
- 국제연합(UN): 2030년까지 데이터센터 100% 재생 에너지 전환 촉구



조선일보

<https://www.chosun.com/weeklybiz/2025/08/28>



AI타임스

<https://www.aitimes.com/news/articleView>

AI 전환이 불러온 전력 대란... 빅테크, 재생에너지 확보 전면전
우후죽순 데이터센터의 그늘... 전기료 올리고 환경 파괴 논란도

[출처: 조선일보, 2025년 8월28일; AI타임즈, 2025년 9월3일]

환경 지속가능성에 대한 고려(2)

- ▶ 구글이 Gemini AI 모델 질의 당 전력 사용량 공개
 - 평균 '중간값 질의' 1건당 0.24 Wh (전자레인지 1초 가동량과 유사)
 - ▶ 전력 소비 구성
 - TPU 칩 58%, CPU·메모리 25%, 백업 장비 10%, 인프라(냉각·전력변환 등) 8% 차지
 - 칩 효율화 단독만으로는 전체 에너지 절감 한계
 - ▶ 에너지 효율 향상
 - 1년간 질의 당 전력 소비 33배 감소
 - 탄소 배출 0.03g, 물 사용량 0.26mL 공개
- ⇒ **AI 에너지 효율성 '라벨링' 필요성 대두**

[출처: MIT Technology Review, 2025년 8월21일]

목차

4. 맺음말

AI의 가능성과 가치

- ▶ AI는 SR을 포함한 HTA의 여러 단계에서 **자동화를 통한 효율성 향상** 가능성을 입증하고 있음
- ▶ 문헌 스크리닝, 데이터 추출, 비뿔림 위험 평가 등 반복적 작업을 신속하게 처리하여 **시간과 자원 제약**을 극복할 수 있음
- ▶ AI의 주요 가치는 **시의성(timeliness)**과 **견고성(robustness)** – 최신 근거를 신속히 반영하고, 프로세스의 일관성과 신뢰성을 강화함
- ▶ 결과적으로, AI는 HTA의 근거 생성 체계에서 **변화 가능성을 가진 중요한 파트너**로 인식되고 있음

한계와 책임 있는 활용

- ▶ AI는 환각(hallucination), 편향(bias), 출처 불명확성 등 기술적 한계를 지님
- ▶ **전문가의 엄격한 감독(human oversight)과 검증 과정** 없이는 신뢰할 수 있는 근거 생산이 어려움
- ▶ AI 활용의 안전성과 투명성을 확보하기 위해 **국제 표준 프레임워크(RAISE 등)** 기반의 거버넌스 구축이 필요함
- ▶ AI 연구자를 대체하는 것이 아니라, **비판적 판단을 지원하는 보조 수단**임을 명확히 인식해야 함

참고 문헌 및 자료(1)

- 조선일보, 2025년 8월28일 보도자료. 우후죽순 데이터센터의 그늘...전기료 올리고 환경 파괴 논란도.
- 한국보건의료연구원. 『의학 문헌고찰에서 AI 챗봇의 활용 가능성 연구』 연구결과보고서. 2025. 5.
- AI타임즈 보도자료, 2025년 9월3일 보도자료. AI 전환이 불러온 전력 대란...빅테크, 재생에너지 확보 전면전.
- Andersen TH, et al. Using Artificial Intelligence Tools as Second Reviewers for Data Extraction in Systematic Reviews: A Performance Comparison of Two AI Tools Against Human Reviewers. Cochrane Evid Synth Methods. 2025;3(4):e70036.
- Cochrane Training. Artificial intelligence (AI) methods in evidence synthesis: learning live webinar series [Internet]. London: Cochrane Training [cited 2025 November 10]. Available from: <https://training.cochrane.org/AI-in-evidence-synthesis-webinars>.
- DESTINY 홈페이지. <https://destiny-evidence.github.io/website/#alive>
- Eisele-Metzger A, et al. Exploring the potential of Claude 2 for risk of bias assessment: Using a large language model to assess randomized controlled trials with RoB 2. Research Synthesis Methods 2025;16(3):491-508.
- Eastaugh CH, et al. Exploring the Role of Artificial Intelligence in Evidence Synthesis: Insights From the CORE Information Retrieval Forum 2025. Cochrane Evid Synth Methods 2025;3(5):e70049.

참고 문헌 및 자료(2)

- Fleurence RL, et al. Generative Artificial Intelligence for Health Technology Assessment: Opportunities, Challenges, and Policy Considerations: An ISPOR Working Group Report. ISPOR REPORT 2025;28(2):175-83.
- Hu B, et al. Enhancing Evidence Synthesis Efficiency: Leveraging Large Language Models and Agentic Workflows for Optimized Literature Screening. Cochrane Evid Synth Methods. 2025;3(6):e70042.
- Kallmes KM, et al. Human-in-the-Loop Artificial Intelligence System for Systematic Literature Review: Methods and Validations for the AutoLit Review Software. Cochrane Evid Synth Methods. 2025;3(6):e70059.
- Lau O, Golder S. Comparison of Elicit AI and Traditional Literature Searching in Evidence Syntheses Using Four Case Studies. Cochrane Evid Synth Methods. 2025;3(6):e70050.
- Lieberum JL, et al. Large language models for conducting systematic reviews: on the rise, but not yet ready for use - a scoping review. J Clin Epidemiol 2025;181:111746.
- MIT Technology Review, 2025년 8월21일 보도자료. In a first, Google has released data on how much energy its new Gemini AI uses.
- Murton M, et al. Harnessing Large-Language Models for Efficient Data Extraction in Systematic Reviews: The Role of Prompt Engineering. Cochrane Evid Synth Methods. 2025;3(6):e70058.

참고 문헌 및 자료(3)

- Rose CJ, et al. Using a Large Language Model (ChatGPT-4o) to Assess the Risk of Bias in Randomized Controlled Trials of Medical Interventions: Interrater Agreement With Human Reviewers. *Cochrane Evid Synth Methods*. 2025;3(5):e70048.
- Rubinstein M, et al. Using GPT-4 for Title and Abstract Screening in a Literature Review of Public Policies: A Feasibility Study. *Cochrane Evid Synth Methods*. 2025;3(3):e70031.
- Simmons Z, et al. Assessing the Feasibility and Acceptability of a Bespoke Large Language Model Pipeline to Extract Data From Different Study Designs for Public Health Evidence Reviews. *Cochrane Evid Synth Methods*. 2025;3(6):e70061.
- Sul AR. Utilization of AI in Conducting Medical Systematic Reviews. *J Health Tech Assess* 2025;13(1):1-7.
- Taneri PE. Human Versus Artificial Intelligence: Comparing Cochrane Authors' and ChatGPT's Risk of Bias Assessments. *Cochrane Evid Synth Methods*. 2025;3(5):e70044.
- Teperikidis L, et al. Using ChatGPT to perform a systematic review; a tutorial. *Minerva Cardiol Angiol* 2024;72:547-567.
- Thomas J, et al. Responsible AI in Evidence Synthesis (RAISE). Draft for Consultation and Revision 3. June 2025. <https://osf.io/fwaud/overview>
- Xing X, et al. Leveraging AI for Meta-Analysis: Evaluating LLMs in Detecting Publication Bias for Next-Generation Evidence Synthesis. *Cochrane Evid Synth Methods*. 2025;3(5):e70047.

경청해 주셔서 감사합니다.



ahramsul@neca.re.kr
ahram.sul@gmail.com