

Article

체계적 문헌고찰에서 생성형 AI 기반 검색전략의 품질 평가와
인간 전문가(의학사서)와의 검색 성능 비교박혜원^{1,†}, 신동원^{2,†}, 김나진³¹인제대학교 의학도서관, ²순천향대학교 부속 서울병원 학술정보팀, ³가톨릭대학교 성의교정 도서관Quality Evaluation of Generative AI-Based Search Strategies in Systematic Reviews
and Comparison of Search Performance with Human Expert (Medical Librarian)Hye Won Park^{1,†}, Dong Won Shin^{2,†}, Na Jin Kim³¹Medical Library, Inje University, Busan, ²Medical Library, Soonchunhyang University Seoul Hospital, Seoul, ³Medical Library, The Catholic University of Korea, Seoul, Korea

Reproducible and sensitive literature search strategies are essential to the methodological rigor of systematic reviews (SRs). Recently, generative artificial intelligence (AI) tools such as ChatGPT and Gemini have been explored as aids for search strategy development; however, concerns remain regarding logical errors, limited understanding of controlled vocabularies, and reproducibility. This study evaluates the quality of search strategies generated by generative AI for a specific clinical topic and compares their retrieval performance with those developed by professional medical librarians.

The research topic was defined as the effectiveness of psychotherapy for patients with obesity. Relevant Cochrane Reviews were identified in the Cochrane Library database using a search strategy incorporating both MeSH terms and text word keywords. Based on the final reference lists of the included reviews, a gold standard comprising a total of 139 studies was constructed. The comparative experiment involved generative AI models (ChatGPT-5, Gemini 2.5) and a human expert (medical librarian). The AI group operated in a zero-shot environment using three levels of prompts based on researcher proficiency, while the human expert formulated strategies adhering to the Cochrane Handbook and PRISMA-S guidelines. The derived search strategy was applied to three major academic databases (PubMed, EMBASE, and the Cochrane Library). Its performance was evaluated qualitatively using the PRESS 2015 Evidence-Based Guideline Checklist and quantitatively by calculating sensitivity (recall) and precision.

As prompt levels advanced, generative AI showed improved structural systematicity in search strategies; however, limitations persisted in the accurate application of database-specific syntax (truncation, field tags) and controlled vocabularies (MeSH/Emtree). In the performance evaluation, the human expert achieved the highest sensitivity (54.7%), followed by Gemini 2.5 (advanced prompt) at 49.6% and ChatGPT-5 (intermediate prompt) at 46.8%. Precision was low (less than 1%) for all groups. The analysis of missing studies revealed that human expert omitted some recent intervention-related literature, such as digital health studies, due to relying on traditional conceptual boundaries of “psychotherapy.”

Generative AI is a useful auxiliary tool for deriving initial search concepts and expanding natural language terms; however, verification by medical librarians remains essential to ensure appropriate controlled vocabulary use and syntactic accuracy. Medical

librarians, in contrast, developed comprehensive and reproducible high-sensitivity search strategies for systematic reviews, though further refinement is required to improve precision. Accordingly, this study proposes a Human-AI Collaboration model that integrates generative AI automation with the methodological expertise of medical librarians, highlighting the continued importance of medical librarians in high-quality evidence synthesis in the AI era. Future research should encompass diverse clinical questions beyond a single topic and continuously evaluate the performance of rapidly evolving AI models. [J Korean Med Libr Assoc 2025;52(1):28-47]

Keywords: Generative AI, Literature Search Strategy, Medical Librarian, PRESS 2015 Guideline, Prompt Engineering

Submitted: December 2, 2025

Revised: December 23, 2025

Accepted: December 26, 2025

Corresponding Author:

Na Jin Kim
Medical Library, The Catholic University
of Korea, 222 Banpo-daero, Seocho-gu,
Seoul 06591, Korea
Tel: +82-2-3147-8164
Fax: +82-2-532-3112
E-mail: kimnj@catholic.ac.kr

*이 저자들은 공동 1저자로 본 연구에
동등하게 기여하였음.

ORCID

박혜원
<https://orcid.org/0000-0001-9374-7848>
신동원
<https://orcid.org/0009-0001-6268-1150>
김나진
<https://orcid.org/0000-0001-7280-9579>

서론

1. 연구의 필요성 및 목적

의학 및 보건의료 분야에서 체계적 문헌고찰(Systematic Review, SR)은 근거기반 의사결정(Evidence-Based Decision-Making)의 핵심적 연구 설계로 자리매김하고 있다[1-3]. SR의 신뢰성과 타당성을 확보하기 위해서는 연구 질문에 적합한 문헌을 포괄적이고 일관된 방식으로 탐색하는 문헌검색 전략(Search Strategy)이 필수적이다. Cochrane Handbook for Systematic Reviews of Interventions(Version 6.5, 2024) 또한 MeSH와 같은 통제어 기반 용어 체계와 자연어 키워드를 병행한 “고민감도(Highly Sensitive) 문헌검색 전략”을 SR의 필수 요건으로 제시하고 있다[1,4,5].

이와 같은 배경 속에서 의학사서(Medical Librarian)의 역할은 근거기반의학(Evidence-Based Medicine)의 발전과 함께 근본적인 변화를 겪어왔다. 전통적으로 자료 제공, 분류, 보존이라는 기술적 업무를 담당하던 의학사서는 자료 검색 및 근거 상담 서비스를 제공하는 정보 전문가로 전환되었고[6], 2000년대 이후에는 체계적 문헌고찰의 문헌검색 전략을 수립하는 정보검색 전문가로 자리매김하였다[7]. 이러한 역할 변화는 Cochrane Handbook for Systematic Reviews of Interventions에서 의학사서 또는 정보검색 전문가를 SR 연구팀의 “필수 인력”으로 규정한 것에서도 확인되며[5], 의학사서는 현재 SR의 질적 수준에 직접적인 영향을 미치는 핵심 역할을 담당하고 있다.

SR에서 의학사서가 수행하는 문헌검색 전략 수립은 핵심 개념의 구조화, 데이터베이스별 색인 체계 분석, 통제어와 자연어의 동시 활용, Boolean 연산자와 필드 제한의 체계적 결합, 그리고 각 데이터베이스별 규칙에 따른 검색식 변환 등 다층적인 전문적 절차로 이루어진다[8]. 이러한 복잡성과 기술적 요건으로 인해 문헌검색 전략의 질은 SR의 결과 해석과 편향 발생 가능성에 직접적 영향을 미치며, 비전문가가 독자적으로는 상당한 난이도를 지닌다. 선행연구에서도 의학사서는 통제어 활용 능력, 필드 제한 검색, Boolean 연산자 구성, 최신 색인어 반영 등에서 비전문가보다 우수한 검색 정확도와 재현성을 보이는 것으로 보고되었다[9].

SR 수행은 상당한 시간과 인적 자원을 요구하는 복합적 절차로 진행된다. 선행 연구에서는 SR의 기획에서 출판까지 평균 5명 이상의 연구팀과 약 67.3주가 필요함을 보고하고 있으며[10], 이 중 의학사서는 연구자 요구 분석, 문헌검색 전략 개발, 데이터베이스별 검색식 변환, 검색 수행 문서화 등의 과정에 평균 27시간 이상을 투입하는 것으로 나타났다[11]. 이는 문헌

검색이 SR 전체 과정에서 매우 높은 노동 강도와 전문성을 필요로 하는 단계임을 보여주며, 의학사서의 전문성과 기여도가 매우 중요하다는 사실을 시사한다.

2. 선행 연구

최근에는 ChatGPT, Gemini, Claude 등 대규모 언어 모델(Large Language Model, LLM) 기반 생성형 AI(Generative Artificial Intelligence)의 발전으로 문헌검색 전략 개발 과정에서도 AI 활용이 시도되고 있다. 일부 연구는 생성형 AI가 검색식 초안 작성, 키워드 브레인스토밍, 연구 질문 구조화(PICO) 등의 초기 단계에서 개념 확장과 속도 측면의 잠재적 장점을 가질 수 있음을 보고하였다[12,13].

그러나 최신 연구들은 생성형 AI가 SR의 표준을 충족하는 독립적 문헌검색 도구로 활용되기에는 뚜렷한 한계가 존재함을 지적하고 있다. 대표적으로, 검색식의 논리적 오류, 최신 통제어 미반영, 색인 체계에 대한 이해 부족, 검색 결과의 재현성 부족, 프롬프트에 따른 출력의 변동성, 할루시네이션 등이 지속적으로 보고되고 있다[14,15]. AI 기반 검색 도구 7종을 평가한 최근 분석에서도 생성형 AI 도구들은 일관성(Consistency), 정확성(Accuracy), 재현 가능성(Reproducibility) 측면에서 SR 기준을 충족하지 못하는 것으로 나타났다[16]. 또한 인간과 AI의 문헌검색 성능을 직접 비교한 최근 연구들 역시 AI의 빠른 정보 수집 능력에도 불구하고, 근거기반 연구에서 요구되는 정확성과 신뢰성 측면에서는 정보검색 전문가의 역할을 대신하기 어렵다는 결과를 보고하고 있다[17]. 최근 문헌에서도 AI는 개념 도출, 초기 검색식 생성, 관련 키워드 추천 등에서 의학사서의 업무 효율을 향상시킬 수 있으나[18,19], 최종 검색식의 질과 재현성 확보를 위해서는 정보검색 전문가의 검증이 필수적이라고 제안하고 있다[12,16,20].

하지만 기존의 선행연구들은 주로 SR에서의 AI 활용 가능 여부 및 포괄적인 관점에서의 비교 분석이 주로 이루어졌으며, AI가 제안한 문헌검색 전략을 관련지침에 근거하여 세부적으로 평가한 연구 자료는 부족한 것으로 확인된다. 예를 들어 통제어의 실제 색인 여부, 사용된 필드의 통일성, 인접 연산자의 작동 여부 등 AI가 제시한 문헌검색 전략의 정확도와 실제적인 활용 가능성을 평가하기 위해서는 각 데이터베이스별 고유한 특성과 기능에 근거하여 AI가 제안한 문헌검색 전략에 대해 구체적으로 세부요소들을 평가할 필요가 있다.

이러한 배경에서 본 연구는 특정 연구 주제를 대상으로 생성형 AI가 제안한 문헌검색 전략을 실제 학술 데이터베이스에 적용하여 그 품질을 검증하고, 인간 전문가(의학사서, 이하 의학사서)가 설계한 문헌검색 전략과 검색 성능을 비교하고자 한다. 이를 통해 생성형 AI의 문헌검색 전략 개발 과정에서의 가능성과 한계를 규명하고, 향후 인간-AI 협업 기반의 문헌검색 전략 설계를 위한 실증적 근거를 마련하고자 한다.

본 연구는 다음 세 가지 구체적 연구 질문을 설정하였다.

1. 프롬프트 수준(초급-고급)에 따라 생성형 AI가 생성하는 문헌검색 전략의 구조적, 내용적 특징은 어떠한가?
2. 생성형 AI가 생성한 검색식은 PRESS(Peer Review of Electronic Search Strategies) 2015 Evidence-Based Guideline Checklist[21]의 기준을 충족하는가?
3. 의학사서와 생성형 AI는 문헌검색 전략 수행 능력에는 차이가 있는가?

본 연구는 생성형 AI를 SR 문헌검색 전략 개발 과정에 어떻게 통합할 수 있는지에 대한 근거를 제시함으로써, 의학사서의 전문적 역할 변화와 인간-AI 협업의 방향성 탐색에 중요한 시사점을 제공할 것으로 기대된다. 특히 관련 지침을 근거로 생성형 AI가 제안한 문헌검색 전략에 존재하는 특징, 부적절한 사용법 및 문제점 등을 평가하고 제시함으로써, 향후 SR연구를 진행하는 연구자와 의학사서 등 관련 전문가들이 AI를 활용한 SR 문헌검색 전략 개발의 실무과정에서 검증하여야 할 요소 등을 확인하고자 한다.

연구방법

1. 연구 설계

본 연구는 SR에서 생성형 AI가 제안한 문헌검색 전략의 품질을 검증하고, 의학사서가 설계한 문헌검색 전략과 검색 성능을 비교하는 연구이다. 이를 위해 Cochrane Library 데이터베이스에 수록된 고품질 체계적 문헌고찰을 대상으로 골드 스탠다드(Gold Standard) 문헌 집합을 먼저 구축하였다. 이후 의학사서와 생성형 AI 도구를 활용한 AI 검색 팀이 동일한 연구 질문을 바탕으로 서로 독립적으로 검색 전략을 수립하고 검색을 수행하도록 하였다.

두 검색 전략의 성능은 골드 스탠다드 문헌 집합을 기준으로 비교하였으며, 민감도(Recall), 정밀도(Precision)를 핵심 성능 지표로 산출하여 상대적 검색 성능과 문헌 누락 양상을 객관적으로 평가하였다. 또한 생성형 AI가 도출한 검색 전략의 질적 수준은 PRESS 2015 Evidence-Based Guideline Checklist[21]를 적용하여 검색식 구성 요소별로 체계적으로 비교, 평가하였다.

2. 연구 주제 선정 및 골드 스탠다드 구축

본 연구는 생성형 인공지능이 SR의 문헌검색 전략을 적절히 생성할 수 있는지를 평가하기 위해 Cochrane Database of Systematic Reviews의 고품질 SR을 대조군으로 사용하였다. 적절한 비교를 위해 다음과 같은 기준에 따라 골드 스탠다드를 선정하였다.

첫째, 보건의료 분야에서 임상적 중요성이 높고 연구자와 독자가 개념을 명확히 이해할 수 있는 주제일 것.

둘째, Cochrane Library에서 해당 주제에 대한 Cochrane Review가 이미 출판되어 있어, 엄격한 방법론적 기준을 통과한 포함 문헌 목록을 기준 집합으로 활용할 수 있을 것.

셋째, Cochrane Database of Systematic Reviews에 게재되어 방법론적 품질관리(Quality Control)가 확보된 문헌일 것.

넷째, 생성형 AI의 학습 데이터 최신성 한계를 고려하여 보편적, 표준화된 임상 주제일 것.

위 기준에 따라 개념적 명확성, 문헌 축적량, 방법론적 품질관리, 그리고 생성형 AI 비교의 공정성을 종합적으로 고려하여 「비만 환자에 대한 심리치료의 효과」를 연구 주제로 선정하였다. Cochrane Library 데이터베이스에 주 키워드에 대한 MeSH 통제어와 자연어 키워드를 결합한 검색식 ([mh "Obesity"] OR "Obesity":ti,ab,kw) AND ([mh "Psychotherapy"] OR

“Psychotherapy”:ti,ab,kw)을 적용하여 연구 주제와 관련된 Cochrane Review를 선별하였다.

본 검색식은 연구 주제의 핵심 개념을 중심으로 각 주요 키워드에 대해 MeSH 통제어휘 ([mh])와 자연어 검색(ti,ab,kw)을 OR 연산자로 결합하고, 두 개념 간에는 AND 연산자를 적용하여 구성하였다.

본 연구에서의 골드 스탠다드는 체계적 문헌고찰 수행을 위한 문헌검색 전략과는 기능적으로 구분된다. 체계적 문헌고찰의 문헌검색 전략은 연구 질문에 부합하는 모든 잠재적 관련 문헌을 최대한 포괄적으로 식별하기 위해 통제어와 자연어를 병행하고, 광범위한 동의어 확장과 다양한 검색 기법을 적용하는 고민감도 전략을 요구한다. 반면, 본 연구의 골드 스탠다드는 서로 다른 검색 전략 간 성능을 동일한 기준 하에서 비교·평가하기 위한 기준 집합을 설정하는데 목적을 둔다.

이를 위해 사전 등록된 프로토콜과 엄격한 방법론적 심사를 거쳐 출판되며, 포함 문헌이 이미 체계적 선별과 품질 관리를 통과한 Cochrane Review를 활용하였다. 본 연구의 검색식은 체계적 문헌고찰을 위한 포괄적 검색 전략을 대체하거나 축약하기 위한 것이 아니라, Cochrane Library의 색인 특성을 고려하여 골드 스탠다드 식별에 필요한 최소 조건을 설정한 것이다. 따라서 본 연구에서 산출된 민감도와 정밀도는 검색 전략의 절대적 성능이 아니라, 동일한 골드 스탠다드를 기준으로 한 상대적 비교 결과로 해석되어야 한다.

골드 스탠다드의 포함 기준은 (1) 비만 환자 대상 연구, (2) 심리치료 중재 연구, (3) Cochrane Review 논문으로 설정하였으며, 제외 기준은 (1) 연구 대상 불일치, (2) 철회 논문, (3) 중복 출판물로 하였다. 검색 결과 총 7건의 Cochrane Review가 확인되었고, 이 중 연구 대상 불일치 2건과 철회 논문 1건을 제외하여 최종 4건을 선정하였다. 선정된 4건의 Cochrane Review에 포함된 1차 연구를 통합하여 총 139편의 문헌을 수집하였다. 서지관리프로그램(EndNote 2025)을 이용한 자동 중복 제거 후 DOI, PMID, 논문 제목, 저자, 저널명, 출판연도를 기준으로 수작업 중복 제거를 수행한 결과, 중복 문헌은 확인되지 않아 최종 139편을 골드 스탠다드로 확정하였다 (Supplementary Materials 1).

3. 학술 데이터베이스 및 생성형 AI 도구 선정

1) 학술 데이터베이스 선정

SR의 핵심 학술 데이터베이스로 PubMed, EMBASE(via Elsevier), Cochrane Library를 선정하였다. 학술 데이터베이스는 보건의료 분야의 대표적인 문헌 데이터베이스로서 SR 수행 시 필수적으로 검색되는 데이터베이스이다[1,5].

2) 생성형 AI 도구 선정

범용적으로 널리 사용되는 생성형 AI 도구 중 ChatGPT-5와 Gemini 2.5를 선정하였다 [22]. 두 가지 도구 모두 유료 버전을 사용하였으며, 사전 학습 데이터나 대화 기록이 없는 새로운 계정을 생성하여 제로샷(Zero-Shot) 환경에서 실험을 수행하였다. 이는 문헌검색 전략 개발이라는 맥락에서 의학사서 개입 없이 단번에 문헌검색 전략을 생성할 수 있는가를 평가하기에 적합한 조건이라고 할 수 있다. 현재 전체 생성형 AI 사용자 중 약 13.5%만이 유료 버전을 사용하고 있다는 조사 결과가 있으나[23], 본 연구에서 생성형 AI의 유료 버전을 선택한 근거는 다음과 같다. 첫째, 본 연구는 생성형 인공지능의 최대 성능을 평가하는 것을 목표로 하

였다. 둘째, 전문적인 학술 연구의 목적으로 생성형 AI를 활용하는 경우 시의성을 반영한 유료 버전을 선택할 가능성이 높다는 현실적 고려를 반영하였다. 실제로 ChatGPT-5 및 Gemini 2.5 Pro는 각 사 공식 발표문에서 향상된 추론 능력, 큰 컨텍스트 창(Context Window), 적은 오류율(Hallucination) 등을 근거로 제시하고 있다[24,25].

4. 프롬프트 설계 및 실행

생성형 AI의 SR에서의 문헌검색 전략 능력을 다층적으로 평가하기 위해 프롬프트를 3단계 수준으로 설계하였다[26,27]. 초급 수준은 학사 수준의 일반 연구자로 상정하여 연구 주제만 제시하는 단순 지시형으로 구성하였다. 중급 수준은 SR을 수행하는 석사 연구자 수준으로 설계하였으며, PICO 개념별 키워드 선정을 요구하였다. 고급 수준은 10년 이상 경력의 박사 연구자 수준으로 상정하여, 통제어(MeSH/Emtree)와 자연어의 병행 사용, 필드 제한 명시, Boolean 연산자의 명확한 구분, PRISMA 보고 지침과 AMSTAR 가이드라인을 준수하는 세부 조건을 제시하였다. 이에 따라 확정된 프롬프트는 다음과 같다(표 1).

2025년 9월 23일에 각 생성형 AI 도구의 기본 검색창에서 순차적 프롬프트 실행을 수행하였다. 구체적으로 초급, 중급 및 고급 프롬프트 각각 3회씩 연속 실행하는 방식으로 진행하여 각 생성형 AI 도구당 총 9회의 프롬프트를 실행하였다(그림 1, 2). 또한 생성형 AI의 출력 변동성을 고려하여 동일 프롬프트를 3회 반복 실행하였으며[28,29], 각 수준별 3회차 실행 결과 중 실제 학술 데이터베이스 검색에 적합한 프롬프트를 대표 검색식으로 선정하였다.

표 1. 확정된 3단계 수준의 프롬프트

수준	내용
초급	I intend to search for literature on psychotherapy for patients with obesity in PubMed, EMBASE via elsevier, and the Cochrane Library, and therefore request the construction of corresponding search strategies for each database.
중급	You are a graduate student in public health and medicine and are currently conducting a systematic review. You intend to search for studies on “psychotherapy for patients with obesity.” Please provide your responses to the following requirements in sequence: <ol style="list-style-type: none"> 1. Construct a table of the research topic using the PICO framework. 2. Summarize the key terms for each concept. 3. Based on the main concepts and conditions, formulate the search strategies for PubMed, EMBASE via elsevier, and the Cochrane Library, respectively.
고급	You are a doctoral researcher in public health and medicine with over ten years of experience in conducting systematic reviews. You intend to search for studies on “psychotherapy for patients with obesity.” Please construct database-specific search strategies under the following conditions: <ol style="list-style-type: none"> 1. Formulate the research topic using the PICO framework (P = patients with obesity, I = psychotherapy) → Apply the P AND I search strategy. 2. For each concept, identify both controlled vocabulary (e.g., MeSH, Emtree) and free-text terms, including synonyms, related terms, broader and narrower terms, as well as academic expressions commonly used in the literature. 3. Present the search strategies in a line-by-line format, explicitly specifying field tags for free-text terms and clearly distinguishing Boolean operators. 4. When applying search techniques (e.g., field tags), adhere to the reporting standards outlined in the PRISMA Statement extensions and AMSTAR guidelines. 5. Finally, provide search strategies appropriate to each database: PubMed, EMBASE via elsevier, and the Cochrane Library

Prompt

You are a graduate student in public health and medicine and are currently conducting a systematic review.

You intend to search for studies on "psychotherapy for patients with obesity."

Please provide your responses to the following requirements in sequence:

1. Construct a table of the research topic using the PICO framework.
2. Summarize the key terms for each concept.
3. Based on the main concepts and conditions, formulate the search strategies for PubMed, Embase, and the Cochrane Library, respectively.

Here's a structured response to your request:

Answer

1. PICO Framework Table

Element	Description
P (Population)	Patients with obesity (adults or children, depending on inclusion criteria)
I (Intervention)	Psychotherapy (e.g., cognitive behavioral therapy, interpersonal therapy, motivational interviewing, etc.)
C (Comparison)	Usual care, lifestyle modification programs, no treatment, pharmacological therapy, or other interventions
O (Outcomes)	Weight reduction, BMI reduction, adherence to treatment, eating behavior changes, psychological outcomes (depression, anxiety, quality of life), metabolic outcomes

그림 1. ChatGPT-5 중급 프롬프트 3회차 PubMed 검색전략 답변 화면

The screenshot displays the Gemini 2.5 web interface. On the left is a sidebar with navigation options like '새 채팅', 'Storybook', and 'Gems 탐색하기'. The main area is split into two panels. The left panel, titled 'Prompt', contains a text input for a user acting as a doctoral researcher in public health and medicine, requesting a systematic review search strategy for 'psychotherapy for patients with obesity' using the PICO framework. The right panel, titled 'Answer', shows the AI's response. It includes an introduction to PubMed's MeSH and syntax, followed by 'Table 3: PubMed Search Strategy'. This table lists six lines of search terms, combining controlled vocabulary (MeSH terms in [tiab] tags) and free-text terms. The search terms are: 1. Overweight OR Obesity; 2. obese [tiab] OR obesity [tiab] OR overweight [tiab] OR adiposity [tiab] OR corpulence [tiab] OR adipose [tiab] OR BMI [tiab] OR "body mass index" [tiab]; 3. #1 OR #2; 4. Psychotherapy; 5. Psychotherapy OR Cognitive Behavioral Therapy OR Psychodrama; 6. psychotherap* [tiab] OR counseling [tiab] OR counselling [tiab] OR CBT [tiab] OR DBT [tiab] OR "behavioral therapy" [tiab] OR "cognitive therapy" [tiab] OR "psychological intervention" [tiab].

그림 2. Gemini 2.5 고급 프롬프트 3회차 PubMed 검색전략 답변 화면

5. 의학사서가 수행한 문헌검색 전략 개발

의학사서 문헌검색 전략은 SR 작성을 위한 체계적 문헌검색 협업을 15년 이상 진행한 의학사서가 수행하였다. 문헌검색 전략 개발은 Cochrane Handbook for Systematic Reviews of Interventions(Version 6.5, 2024), PRESS 2015 Evidence-Based Guideline Checklist, PRISMA-S(2021) 권고사항을 준수하여 진행하였으며, 검색식은 재현성 확보를 위해 문서화하였다(Supplementary Materials 2).

첫째, 선정한 연구 주제 “비만 환자에 대한 심리치료의 효과”에 대해 PICO(Population, Intervention, Comparison, Outcome) 구조에 따라 재정의하고 핵심 개념을 도출하였다. 도출된 핵심 개념에 따라 MeSH/Emtree 통제어를 활용하여 표준화하였다.

둘째, 각 개념에 대한 통제어와 자연어 키워드를 선정하였다. 통제어는 “Obesity”[Mesh], “Overweight”[Mesh], “Psychotherapy”[Mesh], “Cognitive Behavioral Therapy”[Mesh], “Counseling”[Mesh], “Behavior Therapy”[Mesh] 등의 MeSH 통제어를 기본으로 선정하였고, 자연어는 해당 통제어들의 Entry Terms와 Synonyms를 확인하여 관련 용어에 대한 동의어, 약어를 포함하였으며, 해당 주제의 기존 선행 연구들의 제목, 초록, 저자 키워드 등을 확인하여 확정하였다.

셋째, 각 데이터베이스별 검색식 구성요소를 확인하여 선정된 각각의 통제어와 자연어 키워드에 대해 PubMed에서는 [Mesh], [TIAB], [PT] 필드를 사용하였고, EMBASE에서는 /exp, :ti,ab, /it 필드를 사용, Cochrane Library에서는 [mh], :ti,ab 필드를 사용하였다. Boolean 연산자는 개념 내 OR 확장(동의어 집합)과 개념 간 AND 결합 구조로 구성하였으며, 불필요한 제한(Filtering) 검색은 진행하지 않았다.

6. 검색 실행

2025년 10월 12일 20시(한국표준시, KST), 2명의 독립 평가자가 생성형 AI가 작성한 검색식과 의학사서가 개발한 문헌검색 전략을 핵심 학술 데이터베이스 3종에서 동시에 검색 실행하였다. 검색 결과는 서지관리프로그램(EndNote 2025)을 이용하여 저자, 논문 제목, 문헌 유형(Reference Type)을 기준으로 일치하는 논문들은 자동 중복 제거를 수행하였다. 중복 제거가 완료된 문헌은 DOI, PMID, 논문 제목, 저자, 저널명, 출판연도를 기준으로 수동 매칭을 수행하여 골드 스탠다드와의 일치 여부를 확인하였다. 이후 모든 검색 결과에 동일한 절차를 적용하였다.

문헌검색 전략의 성능을 평가하기 위해 혼동행렬(Confusion Matrix) 개념을 적용하였다. 혼동행렬은 검색 결과와 골드 스탠다드 간의 일치 여부를 기반으로 참긍정(True Positive, TP), 거짓긍정(False Positive, FP), 거짓부정(False Negative, FN), 참부정(True Negative, TN)으로 분류하는 방법이다.

본 연구에서 TP는 골드 스탠다드에 포함되며 검색 결과로 회수된 관련 문헌, FN은 골드 스탠다드에는 포함되었으나 검색 결과에서 회수되지 않은 관련 문헌, FP는 검색 결과에는 포함되었으나 골드 스탠다드에는 포함되지 않은 비관련 문헌으로 정의하였다. 체계적 문헌고찰을 위한 문헌검색에서는 전체 비관련 문헌의 총량(TN)을 명확히 정의할 수 없으므로, TN 및 이를 기반으로 한 분류 정확도(Accuracy)는 분석 지표에서 제외하였다. 민감도(Recall)는 골드

스탠다드 문헌 중 검색식으로 회수된 문헌의 비율[TP/(TP+FN)]로, 정밀도(Precision)는 검색 결과 중 골드 스탠다드에 해당하는 문헌의 비율[TP/(TP+FP)]로 산출하였다.

결과

1. ChatGPT-5 분석 결과

1) ChatGPT-5 프롬프트 수준별 검색식 특징

프롬프트 설계 수준에 따라 ChatGPT-5가 생성한 문헌검색 전략의 구조와 구성 요소에는 뚜렷한 차이가 나타났다. 초급 프롬프트에서는 비만과 심리치료 두 개념을 AND 연산자로 단일 라인에서 단순 결합하는 형태가 주로 제시되었고, PubMed와 EMBASE에서는 Title/Abstract, Cochrane Library에서는 Keyword 필드 중심으로 대표 개념 위주의 자연어만 제한적으로 활용되었다. 이 단계에서도 MeSH와 Emtree 등 통제어는 3가지 데이터베이스 모두에서 사용되었으나, PubMed와 Cochrane Library에서 지원되지 않는 단어 중간 절단검색(*)을 적용하는 등 기본 문법 수준의 오류가 관찰되어 연구자의 추가 검증이 필요한 것으로 나타났다.

중급 프롬프트에서는 연구주제를 PICO 구조에 따라 Population과 Intervention을 중심으로 개념화하고, Comparison 및 Outcome까지 정의하는 등 문제 구조화 수준이 향상되었다. 자연어 키워드는 초급에 비해 일부 확장되었고, 세 데이터베이스 모두에서 무작위대조시험(Randomized Controlled Trial) 등 연구디자인 제한 필터를 적용하는 등 문헌검색 전략 요소가 다양해졌다. 그러나 Cochrane Library에서는 통제어 없이 자연어만 제시되거나, PubMed와 Cochrane Library에서 필드 태그를 생략하는 사례가 있었고, MeSH Database에 존재하지 않는 “Interpersonal Therapy”, “Psychodynamic Therapy”를 통제어로 제안하거나 Emtree에서 “overweight”, “psychodynamic therapy” 등을 부정확하게 사용하는 등 색인어 이해의 한계도 드러났다.

고급 프롬프트에서는 검색식이 line-by-line 구조로 제시되고, NEAR와 같은 인접 연산자가 등장하는 등 구조적 복잡성이 가장 높게 나타났다. 자연어 키워드 수도 프롬프트 수준이 높아질수록 비만 관련 용어는 PubMed 기준 초급 2개, 중급 3개, 고급 9개, 심리치료 관련 용어는 초급 7개, 중급 4개, 고급 최대 19개까지 확장되는 등 프롬프트 설계에 따라 개념 확장 용어 다양성이 크게 증가할 수 있음을 확인하였다. 또한 검색식 내 인간 대상 연구(Humans)로 제한하는 내용이 적용되어 있었고, 연구디자인 필터 등의 적용을 선택사항으로 제시함으로써 선택의 폭을 확장시킬 수 있도록 하였다. 그럼에도 불구하고 Cochrane Library에서의 절단검색 오류, 일부 부정확한 통제어 사용 등은 프롬프트 수준이 높아져도 완전히 해소되지 않아, 프롬프트를 정교하게 설계하더라도 최종 문헌검색 전략의 검증과 교정에는 여전히 의학사서의 개입이 필요함을 시사하였다.

2) PRESS 2015 Guideline Evidence-Based Checklist 기반 문헌검색 전략 품질 평가

ChatGPT-5의 프롬프트 수준별 검색식 구성 요소를 객관적으로 평가하기 위해 PRESS 2015 Guideline Evidence-Based Checklist[21]를 활용하여 생성형 AI별 프롬프트 수준별 문헌검색 전략 품질을 평가하였고(Supplementary Materials 3), ChatGPT-5가 제시한 문헌검색 전략을 다음과 같이 확인하였다.

첫째, Boolean and Proximity Operators 영역에서는 초급, 중급 및 고급 프롬프트 모두에서 Boolean 연산자와 괄호를 활용한 Nesting 기법이 적절하게 구현되었다. 주목할만한 사항으로는 고급 프롬프트의 PubMed 전략에서는 인간 대상 연구로 제한하기 위해 NOT 조합이 유일하게 사용되었고, Cochrane Library와 EMBASE 전략에서는 NEAR 근접 연산자가 적용되었다. 둘째, Subject Headings 영역에서는 프롬프트 수준이 높아질수록 통제어의 활용 빈도가 증가했다. 다만 Cochrane Library의 중급 전략에서는 통제어가 사용되지 않아, 프롬프트 설계 단계에서 통제어 사용 지시를 명시할 필요가 있음을 확인했다. 셋째, Text Word Searching 영역에서도 프롬프트 수준이 상향됨에 따라 자연어의 사용 빈도가 늘어났다. 초급과 중급 단계는 대체로 10개 이내였으나, 고급 단계에서는 약 20~30개까지 증가했다. 한편 초급 전략에서 절단검색(*) 연산자의 위치가 부적절하게 제시된 사례가 확인되어, 실제 적용 전 연구자 검토가 필요함이 드러났다. 필드 태그는 전체적으로 ti, ab가 사용되었으며, 데이터 베이스에 따라 pt, kw 등의 추가 필드가 병기되는 경향이 관찰되었다. 넷째, Spelling, Syntax, and Line Numbers 영역에서는 세 수준 모두 철자 오류는 없었고, 고급 프롬프트에서 라인 수가 증가해 구조적 세분화가 이루어진 것이 확인되었다. 다만 일부 자연어 항목에서 절단검색 연산자 적용이 적절하지 못한 위치에서 사용되어 실행 누락 또는 의도와 다른 결과를 야기할 수 있는 검색 라인이 존재했다. 마지막으로 Limits and Filters 영역은 중급 프롬프트에서는 Randomized Controlled Trials 연구디자인으로 결과를 한정하기 위한 필드 태그가 적용되었고, 고급 PubMed 전략에서는 인간 대상 연구 제한이 추가되었다.

3) 학술 데이터베이스 검색 결과

프롬프트 수준에 따라 3개 핵심 데이터베이스(PubMed, EMBASE, Cochrane Library)의 검색 결과 합계는 12,332건(중급)에서 88,997건(고급)까지 약 7.2배 차이를 보였다(표 2, Supplementary Materials 4). 중급 프롬프트는 고급 대비 약 86% 감소된 결과를 생성하여, 프롬프트 수준별로 검색결과가 상당한 차이를 보였다. 이는 중급 프롬프트에서 연구디자인 제한 필터가 적용됨에 따라 전체적인 검색결과에 영향을 주었던 것으로 확인할 수 있었으며, Cochrane Library의 경우 통제어가 사용되지 않았던 사항도 고려할 수 있다.

각 프롬프트 단계별 제시된 문헌검색 전략의 데이터베이스 검색 적합성에 따라 ChatGPT-5의 경우 검색 수행 시 오류가 없고, 실제적인 검색이 가능한 중급 프롬프트의 문헌검색 전략을 대표 검색식으로 선정하였다.

2. Gemini 2.5 분석 결과

1) Gemini 2.5 프롬프트 수준별 검색식 특징

Gemini 2.5는 ChatGPT-5와 유사하게 프롬프트 수준에 따라 문헌검색 전략의 구조와 복

표 2. ChatGPT-5 프롬프트 수준별 검색 결과(건)

프롬프트 수준	PubMed	EMBASE	Cochrane Library	합계
초급 프롬프트	18,588	35,183	6,623	60,394
중급 프롬프트	3,175	5,522	3,635	12,332
고급 프롬프트	23,642	56,604	8,751	88,997

잡성이 체계적으로 변화하였으나, 통제어 활용, 자연어 확장 방식, 절단검색 연산자 사용, 필드 태그 적용 측면에서 차별화된 패턴을 보였다. 통제어의 경우 초급 프롬프트는 Cochrane Library에서 태그 표기 오류('Obesity':MeSH)를 포함해 일부 검색이 정상 실행되지 않았으나, EMBASE와 PubMed에서는 비교적 적절한 MeSH/Emtree 통제어 선정 양상을 보였다. 중급과 고급 프롬프트에서는 Cochrane Library와 EMBASE에서 통제어를 정확히 적용했음에도 불구하고, PubMed에서는 두 수준 모두 통제어를 전혀 사용하지 않아 학술 데이터베이스 특성 이해가 제한됨을 보여주었다.

자연어 키워드 수는 프롬프트 수준별로 뚜렷한 변화를 보였다. 비만 관련 용어는 초급(2개)에서 중급(최대 14개)으로 크게 증가한 뒤 고급에서는 10개로 정제되었고, 심리치료 용어는 중급에서 최대 24개까지 급증했다가 고급에서는 6개로 대폭 축소되었다.

절단검색 연산자(*)의 사용 패턴도 프롬프트 수준 간 차이가 컸다. 초급은 obes, psycho*, behav* 등 광범위한 용어 확장을 시도하였으나, 중급에서는 절단검색 연산자를 전혀 사용하지 않고 명시적 용어 나열 방식으로 전환하였고, 고급은 핵심 개념에만 제한적으로 적용하여 가장 안정적인 패턴을 보였다. 자연어 필드 태그 활용 측면에서 고급 프롬프트는 모든 데이터베이스에서 체계적으로 자연어 필드 제한([tiabl, :ti,ab,kw, :ti,ab])을 적용하였으나, 필터 제한은 모든 프롬프트 수준에서 전혀 사용하지 않은 것이 두드러진 특징이었다.

검색식 구조는 초급이 2~3단계의 단순 결합 구조였던 반면, 중급은 모든 자연어를 단일 라인으로 통합하여 제시해 구조화 수준이 낮았다. 반면 고급 프롬프트는 통제어 선정, 자연어 필드 제한, 통합 단계, 최종 Boolean 연산자 결합 등 총 8단계의 line-by-line 구조를 갖추어 가장 체계적이었다. 고급 단계에서는 비만 심각도(Obesity class 1·2·3) 등 임상적으로 의미 있는 세부 개념도 자연스럽게 포함되었다.

한편 EMBASE 검색식에서 초급·중급 프롬프트 모두 실제 존재하지 않는 Emtree 용어인 'overweight'/exp를 제안하였으나, 실제 검색 결과는 'obesity'/exp와 동일하게 처리되는 것으로 확인되었다. 이는 Gemini 2.5가 색인 체계 이해에 일부 오류를 보이지만, EMBASE 자체에서 자동 매핑이 이루어져 실제 검색에서는 문제가 발생되지 않았다.

Gemini 2.5는 프롬프트 수준 증가에 따라 검색 구조의 체계화와 용어 정제가 이루어지는 경향을 보였으나, PubMed에서의 통제어 미사용, Cochrane Library의 태그 오류, 절단검색 연산자 및 필드 태그 적용의 불일치 등 데이터베이스별 규칙 이해도는 제한적이었다. 이는 고급 프롬프트에서도 여전히 의학사서의 검증 과정이 필수적임을 시사한다.

2) PRESS 2015 Guideline Evidence-Based Checklist 기반 문헌검색 전략 품질 평가

Gemini2.5의 프롬프트 수준별 검색식 구성 요소를 객관적으로 평가하기 위해 PRESS 2015 Guideline Evidence-Based Checklist[21]를 활용하여 생성형 AI별 프롬프트 수준별 문헌검색 전략 품질을 평가하였고(Supplementary Materials 5), Gemini2.5가 제시한 문헌검색 전략에서 프롬프트 수준에 따른 차이가 확인되었다. 첫째, Boolean and Proximity Operators 영역에서는 초급, 중급, 고급 프롬프트 모두 Boolean 연산자의 적절한 활용과 괄호를 이용한 Nesting 기법을 구현하였으나, NOT 연산자 및 근접 연산자는 활용되지 않았다. 둘째, Subject Headings 영역에서는 프롬프트 수준별로 상이한 통제어 적용 패턴이 관찰되었다. 초급 프롬프트는 “비만”, “심리치료” 개념에 대해 상·하위어를 포함한 2~4개의 통제어를 선정하여 민감도 높은 문헌검색 전략을 제시하였으나, Cochrane Library에서 통제어 태그

오류가 발생하였다. 중급과 고급 프롬프트는 PubMed 검색 시 통제어를 활용하지 않았으며, 고급 프롬프트는 EMBASE와 Cochrane Library에서 각 개념의 통제어를 동일하게 사용하여 검색의 일관성을 유지하였다. 셋째, Text Word Searching 영역에서는 중급 프롬프트가 동의어 및 관련 용어를 가장 포괄적으로 확장하였다. 절단검색 연산자는 중급 프롬프트에서 사용하지 않았고, 초급 프롬프트에서는 넓은 범위로(psych*, obes*) 사용되었으며, 고급 프롬프트에서는 핵심 용어 중심으로(psychotherapy*, obesity*) 적용되었다. 필드 태그는 고급 프롬프트만이 데이터베이스별로 적용하였다(PubMed [tiab], EMBASE :ti,ab,kw, Cochrane Library :ti,ab). 넷째, Spelling, Syntax, and Line Numbers 영역에서는 세 프롬프트 모두 철자 및 구문 오류가 없었으며, 검색식 라인 구조는 초급 3라인, 중급 1라인, 고급 8라인으로 나타났다. 마지막으로, Limits and Filters 영역은 모든 프롬프트에서 적용되지 않았다. 고급 프롬프트는 통제어와 자연어의 통합, 필드 태그 활용, line-by-line 구조를 통해 PRESS 2015 Guideline Evidence-Based Checklist 기준을 가장 충실히 준수하였다.

3) 학술 데이터베이스 검색결과

프롬프트 수준에 따라 3개 핵심 학술 데이터베이스(PubMed, EMBASE, Cochrane Library)의 검색 결과 합계는 43,179건(고급)에서 166,629건(중급)까지 약 3.9배 차이를 보였다(표 3, Supplementary Materials 6). 고급 프롬프트는 중급 대비 약 74% 감소된 결과를 생성하여, 정교한 프롬프트가 더 정제된 검색 결과를 도출함을 확인하였다.

초급 프롬프트는 Cochrane Library에서 통제어 태그 기법 오류('Obesity':MeSH 대신 [mh "Obesity"] 필요)로 검색 실행이 불가능하였다. 이는 중급 및 고급 프롬프트에서 해결되어 Gemini2.5가 올바른 MeSH 태그 기법을 학습하였음을 보여주었다.

3. 생성형 AI 검색식 주요 항목 비교

PRESS 2015 Guideline Evidence-Based Checklist[21]에서는 SR 문헌검색 전략의 주요 항목으로 정확한 통제어 활용과 민감도 향상을 위한 충분한 자연어 확장을 제시하고 있다. 이에 본 연구에서는 통제어와 자연어를 중심으로 ChatGPT-5와 Gemini 2.5가 생성한 검색식을 비교하였다(표 4).

통제어 활용 측면에서 ChatGPT-5는 모든 프롬프트 수준에서 통제어를 사용하였으나, 미색인 MeSH 용어와 부정확한 Emtree 용어 선정 오류가 고급 프롬프트까지 지속되었다. Gemini2.5는 중급과 고급 프롬프트 모두 PubMed 검색 시 통제어를 사용하지 않았다. 이러한 특징은 두 생성형 AI 모두 PRESS 2015 Guideline Evidence-Based Checklist의 통제어 활용 기준을 완전히 충족하지 못하는 것으로 나타났다. 자연어 키워드 확장 측면에서는 프롬

표 3. Gemini 2.5 프롬프트 수준별 검색결과(건)

프롬프트 수준	PubMed	EMBASE	Cochrane Library	합계
초급 프롬프트	119,225	31,665	ERROR ^a	N/A
중급 프롬프트	64,384	89,085	13,160	166,629
고급 프롬프트	15,650	22,239	5,290	43,179

^a통제어(MeSH) 태그 기법 생성 오류로 검색 실행 불가.

표 4. ChatGPT-5 vs Gemini 2.5 프롬프트별 PRESS 2015 Guideline Evidence-Based Checklist 주요 항목 비교표

평가 항목	초급 프롬프트로 생성된 검색식		중급 프롬프트로 생성된 검색식		고급 프롬프트로 생성된 검색식	
	ChatGPT-5	Gemini 2.5	ChatGPT-5	Gemini 2.5	ChatGPT-5	Gemini 2.5
정확한 통제어 사용	C 오류	C 오류	E 오류	P 미사용	P, E 오류	P 미사용 E 오류
자연어 사용	2개/7개	2개/9개	3개/4개	12개/24개	9개/19개	10개/6개
통제어와 자연어의 조합	O	O	O	P (통제어 선정 X) E, C (O)	O	P (통제어 선정 X) E, C (O)
학술 데이터베이스별 검색 태그 활용	P, E: (통제어 Tag, 자연어 [tiab]/:ti,ab Tag 사용) C: (통제어 Tag, 자연어 :ti,ab,kw Tag 사용)	P, E: (통제어 Tag, 자연어 all fields Tag 사용) C: (통제어 Tag 오류 있음, 자연어 :ti,ab,kw Tag 사용)	P: (통제어 Tag, 자연어 all fields Tag 사용) E: (통제어 Tag, 자연어 :ab,ti Tag 사용) C: (자연어 all fields Tag 사용)	P: (자연어 all fields Tag 사용) E, C: (통제어 Tag, 자연어 all fields Tag 사용)	P: (통제어 Tag, 자연어 [tiab] Tag 사용) E, C: (통제어 Tag, 자연어 :ti,ab,kw Tag 사용)	P: (자연어 [tiab] Tag 사용) E: (통제어 Tag, :ti,ab,kw Tag 사용) C: (통제어 Tag, :ti,ab Tag 사용)
Boolean 연산자 조합 사용	O	O	O	O	O	O
검색결과(건)	60,394	검색 불가	12,332	166,629	88,997	43,179

P, PubMed; E, EMBASE; C, Cochrane Library.

프트 수준별로 상이한 패턴이 관찰되었다. ChatGPT-5 고급 프롬프트는 심리치료 개념에서 19개의 자연어를 제시하여 점진적 확장 전략을 보였으며, Gemini 2.5 중급 프롬프트는 24개로 최대 확장을 나타냈다. 그러나 Gemini 2.5 고급 프롬프트는 10개로 급격히 축소되어 민감도 저하 가능성이 우려된다. 검색식 구조 측면에서 두 생성형 AI 모두 고급 프롬프트에서 line-by-line 형태의 체계적 구조를 구현하여 재현성을 확보하였다. 그러나 ChatGPT-5는 통제어의 정확성 문제가, Gemini 2.5는 PubMed 통제어 미사용과 불충분한 자연어 선정이 제한점으로 확인되었다.

종합하면 두 생성형 AI의 검색식은 PRESS 2015 Guideline Evidence-Based Checklist의 주요 기준을 부분적으로 충족하였으나, 통제어 정확성과 자연어 확장의 적정성 측면에서 개선이 필요하며, 전문가 검증 없이 단독으로 SR 문헌검색 전략을 구성하기에는 아직 한계가 있는 것으로 평가된다.

4. 의학사서 검색 결과

의학사서가 구축한 핵심 데이터베이스 문헌검색 전략은 연구자들의 독립적 검토를 통해, SR 수행에 요구되는 민감도를 충분히 확보했으며, 통제어와 자연어의 적절한 조합, 그리고 데이터베이스 간 일관된 검색 구조를 유지한 것으로 확인되었다.

최종 검색 결과는 PubMed 19,930건, EMBASE 41,304건, Cochrane Library 6,776건이었으며, 중복 제거 후 총 50,236건이 확인되었다. 이는 ‘비만’과 ‘심리치료’라는 두 개념의 조합을 적용한 검색에서 핵심 데이터베이스 전반에 걸쳐 충분히 높은 민감도를 확보한 결과로, SR 검색에서 요구되는 “초정밀 검색(High Precision)”보다는 “저정밀·고민감도(High Recall, Low Precision) 검색” 전략인 누락을 최소화하는 접근에 적합한 검색 수행이 이루어졌음을

보여주었다.

의학사서가 수행한 문헌검색 전략은 검색 결과의 과포괄(Over-Retrieval) 가능성을 인지하면서도 SR 수행의 목적에 맞게 “누락 최소화(Missing Studies 최소화)”를 우선하는 접근을 유지하였다는 점에서 SR 방법론의 국제 표준을 충족한다.

ChatGPT-5, Gemini 2.5 및 의학사서 검색식의 학술 데이터베이스별 검색 결과는 표 5에서 확인할 수 있다.

5. 검색결과 성능 분석

앞서 선정된 ChatGPT-5 중급 프롬프트 검색결과(12,332건)와 Gemini 2.5 고급 프롬프트 검색결과(43,179건) 그리고 의학사서 검색결과(68,010건) 검색 성능을 평가하기 위해, 생성형 AI가 제시한 검색식과 의학사서의 검색식을 통한 검색 결과를 골드 스탠다드 139건과 비교하여 문헌검색 전략으로 1) 실제 찾아낸 골드 스탠다드에 포함된 논문 수인 TP, 2) 문헌검색 전략으로는 검색되지 않았으나, 실제로는 골드 스탠다드에 포함되어 있어 반드시 검색되었어야 하는 논문 수인 FN, 3) 문헌검색 전략으로 검색되었으나, 실제 연구 주제와 무관하여 골드 스탠다드 포함 대상이 아닌 불필요한 논문 수인 FP, 4) 그에 따른 민감도(Recall), 5) 정밀도(Precision)를 산출하였으며, 그 결과는 다음과 같다(표 6).

ChatGPT-5 중급 프롬프트 검색식을 3개 학술 데이터베이스에 적용한 결과, PubMed 3,175건, EMBASE(via Elsevier) 5,522건, Cochrane Library 3,635건이었으며, 총 12,332건으로 확인되었고, 서지관리프로그램(EndNote 2025)을 활용하여 중복 문헌 4,480건을 제거해 총 7,852건이 최종 분석 대상에 포함되었다. 7,852건의 검색결과와 골드 스탠다드로 선정한 139건의 문헌을 비교한 결과 65건이 포함된 것으로 확인되었다. 민감도(Recall)는

표 5. ChatGPT-5, Gemini 2.5 및 의학사서 검색식의 학술 데이터베이스별 검색 결과(중복 제거 전)

구분	ChatGPT-5(건)	Gemini 2.5(건)	의학사서 검색식(건)
초급 프롬프트	60,394	ERROR	-
중급 프롬프트	12,332	166,629	-
고급 프롬프트	88,997	43,179	68,010

표 6. ChatGPT-5, Gemini 2.5 및 의학사서 검색식의 성능 비교

구분	ChatGPT 5	Gemini 2.5	의학사서 검색식
TP ^a	65건	69건	76건
FN ^b	74건	70건	63건
FP ^c	7,787건	33,102건	50,160건
민감도(Recall) ^d	46.8%	49.6%	54.7%
정밀도(Precision) ^e	0.83%	0.21%	0.15%

^aTrue Positives: 골드 스탠다드에 포함되며 검색 결과로 회수된 관련 문헌.

^bFalse Negatives: 골드 스탠다드에는 포함되었으나 검색 결과에서 회수되지 않은 관련 문헌.

^cFalse Positives: 검색 결과에는 포함되었으나 골드 스탠다드에는 포함되지 않은 비관련 문헌.

^d골드 스탠다드 문헌 중 검색식으로 회수된 문헌의 비율($TP / (TP + FN)$).

^e검색 결과 중 골드 스탠다드에 해당하는 문헌의 비율($TP / (TP + FP)$).

46.8%, 정밀도(Precision)는 0.83%로 산출되었다.

Gemini 2.5 고급 프롬프트 검색식을 3개 학술 데이터베이스에 적용한 결과, PubMed 15,650건, EMBASE(via Elsevier) 22,239건, Cochrane Library 5,290건으로 총 43,179건이 검색되었다. 서지관리프로그램(EndNote 2025)을 활용하여 중복 문헌 10,008건을 제거한 후 최종적으로 33,171건이 분석 대상에 포함되었다. 골드 스탠다드로 선정한 139건 논문 중 69건이 검색결과에 포함된 것으로 확인되었고, 민감도(Recall)는 49.6%, 정밀도(Precision)는 0.21%로 매우 낮게 나타났다.

의학사서가 구축한 검색식을 3개 학술 데이터베이스에 적용한 결과, PubMed 19,930건, EMBASE(via Elsevier) 41,304건, Cochrane Library 6,776건으로 총 68,010건이 검색되었다. 서지관리프로그램(EndNote 2025)을 활용하여 중복 문헌 17,774건을 제거한 후 최종적으로 50,236건이 분석 대상에 포함되었다. 민감도(Recall)는 54.7%, 정밀도(Precision)는 0.15%로 매우 낮게 나타났다.

분석 결과 동일한 연구 주제와 동일한 골드 스탠다드(139편)를 기준으로 생성형 AI(Chat GPT-5, Gemini 2.5)와 의학사서가 설계한 문헌검색 전략의 성능을 비교한 결과 의학사서의 문헌검색 전략은 TP(76편) 기준으로 가장 높은 민감도(54.7%)를 보였으며, ChatGPT-5 (46.8%), Gemini 2.5(49.6%)보다 안정적으로 골드 스탠다드 문헌을 포착하는 것으로 나타났다(그림 3). 그러나 정밀도(Precision)의 경우 SR 수행을 위한 포괄적인 문헌검색 전략인 관련 문헌의 누락 최소화를 위한 목표로 설계되어 모든 검색식의 검색 결과가 매우 낮게 나타났다. 그 중 의학사서의 검색식에서 정밀도가 가장 낮게 나타난 것은 체계적 문헌고찰 수행을 위해 관련 문헌의 누락을 최소화하는 것을 최우선 목표로 설정한 검색 전략의 설계 특성을 반영한 결과로 해석된다(그림 3).

이에 따른 정밀도 저하는 SR 과정에서 연구자의 선별(Screening) 부담을 크게 증가시킬 수 있는 한계로 지적될 수 있어 의학사서 검색식의 누락 연구(Missing Studies)분석을 진행하여 그 이유를 체계적으로 검토하였다.

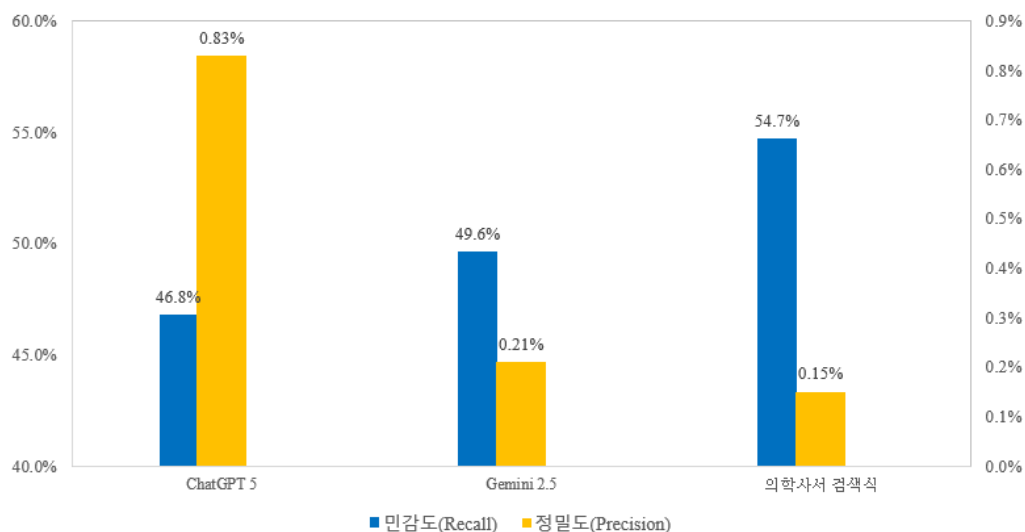


그림 3. ChatGPT-5, Gemini 2.5 및 인간 전문가(의학사서) 검색식의 민감도(Recall), 정밀도(Precision) 비교

6. 누락 연구(Missing Studies) 분석

의학사서 검색식을 통한 핵심 데이터베이스 검색으로는 골드 스탠다드 논문 중 63편이 누락인 것으로 확인되었다. 누락 논문 63편에 대해 문헌검색 전략 주요 관점인 통제어, 자연어, 개념 조합으로 구분하여 누락 원인을 분류하여 살펴보았다.

첫째, 통제어의 경우 비만 관련 통제어는 63편 중 61편(97%)의 제목, 초록, 키워드에서 관련 용어가 분명히 나타났지만, 심리치료와 정신치료 관련 용어의 경우 63편 중 전통적인 심리치료 관련 어휘가 명시된 논문은 사실상 0편에 가까웠으며, 약 10편의 논문에서 행동, 심리와 같은 키워드는 확인되었지만 통제어나 제목에 ‘Psychotherapy’로 표기되지는 않는 것으로 나타났다. 이를 통해 심리치료 개념이 실제 골드 스탠다드에 비해 너무 좁게 정신치료 중심으로 정의되어 있고, 생활습관 및 행동 치료들에 대한 키워드는 빠진 구조라고 확인할 수 있었다.

둘째, 자연어 키워드의 경우도 통제어 문제와 같은 맥락으로 제목과 초록에 자주 등장하는 “Behavioral lifestyle program”, “Mobile health intervention” 자연어 표현이 검색식 내 거의 포함되어 있지 않아 조합 검색 이후 누락된 것으로 확인되었다. 특히 스마트폰, 앱, 온라인에 대한 중재 논문의 경우 자연어 키워드에 디지털 헬스 관련 키워드 세트가 전혀 없어 검색 결과에 포함되지 않은 것으로 파악되었다.

셋째, 위와 같이 Intervention에 해당되는 키워드에 대해 복합 생활습관 및 행동 중재에 대한 키워드가 선정되지 않고 전통적 의미의 심리치료 키워드만이 선정 후 조합되어 개념의 격차 발생으로 인해 골드 스탠다드에 포함되지 않은 논문이 다수 발생한 것으로 확인되었고, 이는 연구 질문과 문헌검색 전략의 불일치로 인한 낮은 민감도의 전형적인 패턴으로 볼 수 있다.

논의

본 연구는 동일한 연구주제와 동일한 골드 스탠다드(139편)를 기반으로 생성형 AI(ChatGPT-5, Gemini 2.5)가 제시한 검색식을 PRESS 2015 Guideline Evidence-Based Checklist[21] 기준에 따라 평가하고, 의학사서가 설계한 문헌검색전략과 검색 성능을 비교하였다. 그 결과, 의학사서의 검색전략은 생성형 AI에 비해 높은 민감도(Recall)를 보여 골드 스탠다드 문헌을 더 많이 포착하였다. 이는 SR에서 요구되는 고민감도 검색전략 특성상 의학사서의 문헌검색 전략이 SR 수행을 위한 체계적 문헌검색의 핵심 요건인 포괄성과 재현성을 충족한 것을 시사한다.

생성형 AI(ChatGPT-5, Gemini 2.5)는 SR을 위한 문헌검색 전략 수립에 일정 수준 기여할 수 있음에도 불구하고, 단독 도구로 활용하기에는 검색 성능과 구조적 완성도가 충분하지 않은 것으로 나타났다. 이는 생성형 AI 기반 문헌검색 전략이 기존의 SR 수준의 재현성과 신뢰성을 확보하지 못한다는 선행연구의 지적과 맥락을 같이한다[30]. PRESS 2015 Guideline Evidence-Based Checklist 항목 비교에서도 여러 요소가 “부분 충족” 수준에 머물러 검색식의 구조적 완성도가 미흡하였고, 특히 통제어 선정, 데이터베이스별 필드 제한, Boolean 연산자 활용 등 핵심 요소를 일관되지 반영하지 못한 점이 확인되어 의학사서의 개입 필요성을 뒷받침한다[21].

세부 오류 양상을 살펴보면, 첫째, PubMed와 Cochrane Library에서 허용되지 않는 중간절단검색 연산자(*) 사용 등 데이터베이스 규칙을 위반한 구문이 다수 관찰되어, 실제 검색 시 의도와 다른 결과를 초래하거나 재현이 불가능해질 위험이 확인되었다. 둘째, MeSH ‘Psy-

chodynamic Therapy', Emtree 'overweight'와 같이 현 시점에서 실제 색인어로 존재하지 않거나 다른 용어로 대체된 통제어를 제안함으로써, 검색 오류 및 편향을 유발할 가능성이 있었다. 셋째, 동일한 주제에 대해 초급-중급-고급 3단계 프롬프트를 9회 반복했음에도, 제안된 문헌검색 전략의 핵심 키워드, 필드 지정, 필터 설정이 매회 달라지는 등 일관성과 재현 가능성이 부족하였다. 넷째, 학습 데이터 시점의 한계로 인해 최신 개념이나 용어 동향이 충분히 반영되지 못할 위험이 존재하며[31], 이는 SR에서 특히 중요한 최신 근거가 누락될 위험성이 있다.

그럼에도 생성형 AI는 문헌검색 전략 수립 전체 과정이 아닌 일부 단계에서 보조 도구로서의 가치가 확인되었다. 연구에서 생성형 AI는 연구 질문을 PICO 구조로 재구성하고, 자연어 기반 관련 개념과 키워드를 폭넓게 제시하는 데에는 비교적 안정적인 성능을 보여, 연구자가 초기 검색식 초안을 구상하고 핵심 개념의 범위를 가늠하는 데 도움이 되었다. 이는 선행연구에서 제안된 것처럼, 의학사서의 검토와 참고문헌 대조 절차가 병행될 경우, 생성형 AI가 문헌검색 전략 개발 과정의 효율성을 높이는 보조 도구로 활용될 수 있음을 시사한다[32].

따라서 생성형 AI는 SR 검색전략 구축 과정에서 의학사서를 대체하기보다는, 전문성을 기반으로 한 Human-in-the-loop 모델 속에서 활용될 때 가장 효과적이다. 즉, 문헌검색 전략의 설계, 검증, 최종 확정 단계에서는 여전히 의학사서의 판단과 책임이 핵심적이라고 할 수 있다.

결론

의학사서와 임상사서의 정보검색 성과를 비교한 선행연구에서 확인한 사항과 같이 의학사서는 구조화된 검색전략, 통제어 및 검색필드의 사용 등 검색에 있어 다양한 기법을 각각의 학술 데이터베이스에 맞게 설정하고 적용하여 일관된 검색을 수행함으로써 품질을 보장한다[9]. 본 연구에서도 생성형 AI가 보여준 불완전한 여러 한계를 통해 의학사서의 개입과 추가적인 확인이 필수적인 요소임을 확인할 수 있었다.

따라서 학술정보검색을 전문적으로 진행하는 의학사서는 생성형 AI를 초기 개념 도출, 키워드 확장, 검색식 초안 작성 등 보조적 단계에서 전략적으로 활용하되, 검색 정확성과 민감도, 재현성을 보장하는 핵심 전문가로서의 역할을 지속해야 한다. 본 연구는 향후 의학도서관 현장에서 인간-AI 협업 기반의 문헌검색 전략 모델을 정립하는데 실증적인 근거를 제공하며, 생성형 AI를 활용한 검색 자동화의 가능성과 한계를 균형있게 이해하는데 중요한 기초 자료가 될 것이다.

한계

본 연구는 연구 대상 주제가 “비만 환자를 위한 심리치료”라는 하나의 임상질문으로 한정되어 생성형 AI의 프롬프트 설계의 다양성이 제한되었다는 점, 그리고 빠르게 진화하는 생성형 AI 모델의 성능 변화를 실시간으로 반영하기 어렵다는 점에서 한계를 가진다. 또한 의학사서의 검색식 설계에서도 범위가 넓고 구체화되지 않은 주제로 인해 적절한 키워드가 제대로 선정될 수 없고, 민감도 높은 검색식을 생성할 수밖에 없었다. 이에 향후 다양하고 구체적인 주제 선정과 핵심 데이터베이스를 포함한 확장 분석, 생성형 AI 모델별 프롬프트 최적화 연구, 인간-AI 협업 기반 검색의 재현성과 품질을 주기적으로 평가하는 후속 연구가 이루어져야 할

필요가 있다.

생성형 AI 사용 명시

원고 작성 과정에서 영문 번역 및 문장 교정을 위해 ChatGPT 5.1과 Gemini 3 Pro를 사용했으며, 저자는 최종 결과물을 검토하였고, 논문의 내용에 대해 전적인 책임을 집니다.

SUPPLEMENTARY MATERIALS

Supplementary materials are only available online from: <https://doi.org/10.69528/jkmla.2025.52.1.28>

이해관계

이해관계 없음.

REFERENCES

1. Kim S, Park D, Seo H, Shin S, Lee S, Lee M, et al. Health technology assessment methodology: systematic review. Seoul: National Evidence-based Healthcare Collaborating Agency; 2020.
2. Paul M, Leibovici L. Systematic review or meta-analysis? Their place in the evidence hierarchy. Clin Microbiol Infect. 2014;20(2):97-100. <https://doi.org/10.1111/1469-0691.12489>
3. Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. BMJ. 1996;312(7023):71-72. <https://doi.org/10.1136/bmj.312.7023.71>
4. Franco JVA, Garrote VL, Escobar Liquitay CM, Vietto V. Identification of problems in search strategies in Cochrane Reviews. Res Synth Methods. 2018;9(3):408-416. <https://doi.org/10.1002/jrsm.1302>
5. Higgins J, Thomas J. Cochrane Handbook for Systematic Reviews of Interventions (current version) [Internet]. 2024 [cited 2025 Nov 15]. Available from: <https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current>
6. Cooper ID, Crum JA. New activities and changing roles of health sciences librarians: a systematic review, 1990-2012. J Med Libr Assoc. 2013;101(4):268-277. <https://doi.org/10.3163/1536-5050.101.4.008>
7. Harris MR. The librarian's roles in the systematic review process: a case study. J Med Libr Assoc. 2005; 93(1):81-87.
8. Lee SJ, Kim NJ. Analysis of the search strategies in Korean authors' systematic reviews to expand the untact search services of medical librarians. J Korean Med Libr Assoc. 2021;48(1_2):1-23. https://doi.org/10.69528/jkmla.2021.48.1_2.1
9. Dadashi A, Zarea Gavgani V, Hajebrahimi S, Mahami-Oskouei M. Comparing the performance of librarians and medical specialists in retrieving clinical evidence: an observational study. Med Ref Serv Q. 2025;44(2):169-186. <https://doi.org/10.1080/02763869.2025.2471886>
10. Borah R, Brown AW, Capers PL, Kaiser KA. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. BMJ Open. 2017;7(2):e012545. <https://doi.org/10.1136/bmjopen-2016-012545>
11. Bullers K, Howard AM, Hanson A, Kearns WD, Orriola JJ, Polo RL, et al. It takes longer than you think: librarian time spent on systematic review tasks. J Med Libr Assoc. 2018;106(2):198-207. <https://doi.org/10.1136/bmjopen-2016-012545>

- org/10.5195/jmla.2018.323
12. Luo X, Chen F, Zhu D, Wang L, Wang Z, Liu H, et al. Potential roles of large language models in the production of systematic reviews and meta-analyses. *J Med Internet Res*. 2024;26:e56780. <https://doi.org/10.2196/56780>
 13. Scherbakov D, Hubig N, Jansari V, Bakumenko A, Lenert LA. The emergence of large language models as tools in literature reviews: a large language model-assisted systematic review. *J Am Med Inform Assoc*. 2025;32(6):1071-1086. <https://doi.org/10.1093/jamia/ocaf063>
 14. Gwon YN, Kim JH, Chung HS, Jung EJ, Chun J, Lee S, et al. The use of generative AI for scientific literature searches for systematic reviews: ChatGPT and Microsoft Bing AI performance evaluation. *JMIR Med Inform*. 2024;12:e51187. <https://doi.org/10.2196/51187>
 15. Bernard N, Sagawa Y Jr, Bier N, Lihoreau T, Pazart L, Tannou T. Using artificial intelligence for systematic review: the example of elicitor. *BMC Med Res Methodol*. 2025;25:75. <https://doi.org/10.1186/s12874-025-02528-y>
 16. Park SG. AI and systematic reviews: can AI tools replace librarians in the systematic search process? *Sci Technol Libr*. 2025;1-22. <https://doi.org/10.1080/0194262X.2025.2521519>
 17. Mostafapour M, Fortier JH, Pacheco K, Murray H, Garber G. Evaluating literature reviews conducted by humans versus ChatGPT: comparative study. *JMIR AI*. 2024;3:e56537. <https://doi.org/10.2196/56537>
 18. Adam GP, DeYoung J, Paul A, Saldanha IJ, Balk EM, Trikalinos TA, et al. Literature search sandbox: a large language model that generates search queries for systematic reviews. *JAMIA Open*. 2024;7(3):o0ae098. <https://doi.org/10.1093/jamiaopen/o0ae098>
 19. Li Y, Datta S, Rastegar-Mojarad M, Lee K, Paek H, Glasgow J, et al. Enhancing systematic literature reviews with generative artificial intelligence: development, applications, and performance evaluation. *J Am Med Inform Assoc*. 2025;32(4):616-625. <https://doi.org/10.1093/jamia/ocaf030>
 20. Alshami A, Elsayed M, Ali E, Eltoukhy AEE, Zayed T. Harnessing the power of ChatGPT for automating systematic review process: methodology, case study, limitations, and future directions. *Systems*. 2023;11(7):351. <https://doi.org/10.3390/systems11070351>
 21. McGowan J, Sampson M, Salzwedel DM, Cogo E, Foerster V, Lefebvre C. PRESS peer review of electronic search strategies: 2015 guideline statement. *J Clin Epidemiol*. 2016;75:40-46. <https://doi.org/10.1016/j.jclinepi.2016.01.021>
 22. Simon F, Nielsen RK, Fletcher R. Public awareness and use of generative AI. *Generative AI and news report 2025: how people think about AI's role in journalism and society*. Oxford: Reuters Institute for the Study of Journalism; 2025. p. 14.
 23. Shin HH. Analysis of generative AI service usage by age group. *KISDI STAT Rep*. 2025;25:1-6. Korean.
 24. OpenAI. Introducing GPT-5 [Internet]. 2025 [cited 2025 Nov 29]. Available from: <https://openai.com/ko-KR/index/introducing-gpt-5/?utm>
 25. Google Cloud. Gemini 2.5 Pro [Internet]. 2025 [cited 2025 Nov 29]. Available from: <https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-pro?utm>
 26. Bourgeois JP, Ellingson H. Ability of ChatGPT to generate systematic review search strategies compared to a published search strategy. *Med Ref Serv Q*. 2025;44(3):279-291. <https://doi.org/10.1080/02763869.2025.2537075>
 27. Wang S, Scells H, Koopman B, Zuccon G. Can ChatGPT write a good Boolean query for systematic review literature search? *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*; 2023 Jul 23-27; Taipei, Taiwan. New York: Association for Computing Machinery; 2023. p. 1426-1436.
 28. Funk PF, Hoch CC, Knoedler S, Knoedler L, Cotozana S, Sofo G, et al. ChatGPT's response consistency: a study on repeated queries of medical examination questions. *Eur J Investig Health Psychol Educ*. 2024;14(3):657-668. <https://doi.org/10.3390/ejihpe14030043>
 29. Liao N, Li C, Gradishar WJ, Klimberg VS, Roshal JA, Yuan T, et al. Accuracy and reproducibility of ChatGPT responses to breast cancer tumor board patients. *JCO Clin Cancer Inform*. 2025;9:e2500001.

- <https://doi.org/10.1200/cci-25-00001>
30. Sul AR. Utilization of AI in conducting medical systematic reviews. *J Health Tech Assess.* 2025;13(1):1-7. Korean. <https://doi.org/10.34161/johta.2025.13.1.001>
 31. OpenAI. Model release notes [Internet]. 2025 [cited 2025 Nov 29]. Available from: <https://help.openai.com/en/articles/9624314-model-release-notes>
 32. Alaniz L, Vu C, Pfaff MJ. The utility of artificial intelligence for systematic reviews and boolean query formulation and translation. *Plast Reconstr Surg Glob Open.* 2023;11(10):e5339. <https://doi.org/10.1097/GOX.0000000000005339>