

# 근거중심의학과 체계적 고찰

## Evidence-based Medicine and Systematic Review

- 김수영 -

교신저자 김수영

한림의대 가정의학과

■ hallymf@gmail.com

근거중심의학(Evidence Based Medicine)은 ‘최신의 근거를 공정하고, 명백하고 현명하게 사용하여 개개의 환자에 대한 의사결정을 하는 것’으로 정의할 수 있으며(1) 근거중심의학의 발전에는 체계적 분석(systematic review) 기법 개발에 힘입은바 크다고 할 수 있다. 근거 중심 의학에서 강조하고 있는 것이 결국 의사결정(decision making)이라고 할 때 그러한 의사 결정에 필요한 근거는 명백한 연구 질문을 통하여 얻어야 한다. 하지만 의학문헌이 폭발적으로 많아지면서 이들 근거들을 효과적으로 결합하여 일정한 결론을 내리는 문제가 중요한 이슈로 대두되고 있다.

따라서 해당 질문에 대한 문헌을 모두 검토해서 일정한 결론을 내리는 고찰(review)의 중요성이 커지고 있다. 이 글에서는 근거중심의학의 꽃이라고 할 수 있는 체계적 고찰의 개념적인 문제, 현황, 체계적 고찰 쓰기 등에 대해 기술하고자 한다.

### I. 개념

고찰은 크게 기술적 고찰(narrative review)과 체계적 고찰로 구별할 수 있다. 기술적 고찰은 현재까지도 고찰의 주류를 이루고 있는 서술 방식이며 해당 분야에 대해 폭 넓은 지식을

제공한다는 장점이 있다. 하지만 기술적 고찰은 인용 혹은 고찰한 문헌이 불완전하다는 문제, 고찰에 이용한 문헌을 선택할 때 주관적으로 한다는 것 그리고 문헌을 분석할 때 과학적이지 않은 방법론을 사용한다는 비판이 있어 왔다. 체계적 고찰은 그러한 기술적 고찰이 가지고 있는 문제점을 보완한 것으로 명확히 정의된 문헌검색에 기반하고, 결정된 기준(criteria)을 이용하여 선택된 문헌의 질을 평가하며, 타당성이 입증된 방법으로 분석을 시행한다는 특징이 있다(2,3)(표 1).

표 1. 비체계적 고찰과 체계적 고찰

특징	비체계적 고찰	체계적 고찰
연구 질문	대부분 광범위	대부분 광범위
검색 전략과 데이터베이스	검색 데이터베이스와 전략이 제공되지 않음	검색 데이터베이스와 전략이 제공되지 않음
연구 선택	특이적이지 않고 자의적일 가능성 있음	기준에 기초한 선택
연구의 질	평가한다고 해도 비정규적 평가	자료 추출 과정에 포함, 공식적으로 평가
논문 검토	변이가 크다	자료 추출표를 통한 비판적 추출
자료 결합	질적 결합	가능하면 양적 결합(메타분석) 그렇지 않으면 질적 결합
추론	때때로 근거 중심	대부분 근거 중심

메타분석(meta analysis)은 같은 질문에 답하는 여러 연구의 결과를 종합적인 측정치로 결합하는 통계적 기법으로 체계적인 고찰과 동의어는 아니다. 메타분석 또한 근거중심의학의 주요한 방법론으로 계량적인 종합수치를 제시함으로써 명확한 의사결정을 가능하게 한다. 비슷하지만 조금 다른 연구로 Pooled analysis(Individual patient data meta analysis)라는 것이 있다. 이것은 연구를 결합할 때 각 연구에 나와 있는 정보를 통해 결합하는 것이 아니고 개별 연구의 실제 데이터를 모아서 이를 통해 분석하는 것을 말한다(2).

체계적 고찰 혹은 메타분석이 기술적 고찰이 가지고 있는 모든 문제를 해결할 수 있지는 않다. 메타분석이 수행되었다고 하더라도 ①질이 나쁜 연구가 포함되거나 질 문제를 무시한 경우, ②이질성에 대해 적절한 주의를 기울이지 못할 때, ③ 무차별적인 자료 병합으로 결론이 정확하지 않은 경우, ④ 출판 비फल, 시간 지체 비फल, 이중 출판 비फल, 언어 비फल, 결과 보고 비फल과 같은 보고 비फल이 문제가 되는 경우에는 메타분석 혹은 체계적 고찰이 잘못된 결론에 도달할 수 있다.

## II. 현황

현재 전 세계적으로 체계적 고찰을 통한 연구 활동이 활발하게 진행되고 있다. 일례로 의학 문헌에 대한 데이터베이스인 PubMed(<http://www.ncbi.nlm.nih.gov/PubMed/>)에서 체계적 고찰(메타분석)의 수를 연도 별로 조사하였을 때 결과는 그림 1과 같다. 체계적 고찰을 이용한 연구는 1990년 이후 급격히 증가하고 있으며 특히 2000년 이후에 활발히 연구가 이루어지고 있다.

전 세계적으로 체계적 고찰 연구 영역을 주도하고 있는 곳은 코크란 연합 (Cochrane Collaboration) (<http://www.cochrane.org>)이다. 이 단체는 영국 역학자 Archi Cochrane이 주창하여 만들어 졌으며, 1980년대 만들어진 Oxford Database of Perinatal Trials가 모체라고 할 수 있다. 1992년 영국에서 처음으로 코크란 센터가 설립되고, 1993년 collaboration이 구성되었다. 코크란 라이브러리를 통해 동시에 검색가능한 데이터베이스는 아래와 같다.

- ① The Cochrane Database of Systematic Reviews(Cochrane reviews)

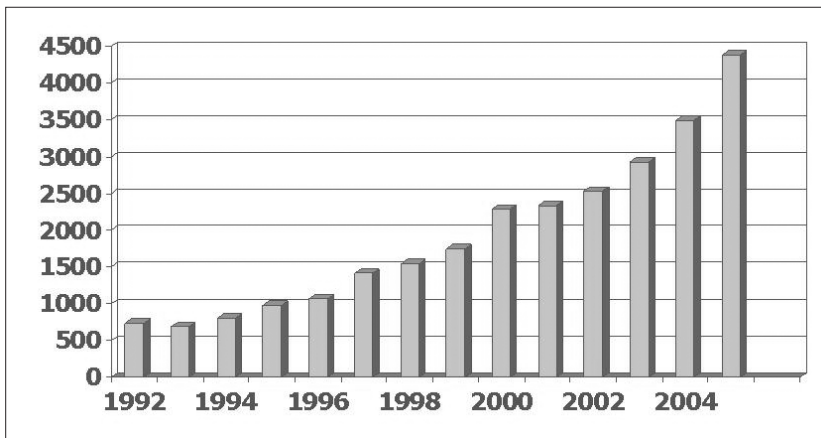


그림 1. Pubmed에 색인된 연도별 체계적 고찰의 수  
 검색식 = meta-analysis[mh] OR meta-analysis[tw] OR systematic review[tw]

- ② The Cochrane Controlled Trials Register  
(Clinical trials)
- ③ The Cochrane Methodology Register  
(Methods studies)
- ④ Cochrane Database of Methodology Reviews  
(CDMR, Methods reviews)
- ⑤ Database of Abstracts of reviews of  
Effectiveness(DARE, Other reviews)
- ⑥ Health Technology Assessment Database
- ⑦ About the Cochrane collaboration and the Cochrane  
review groups
- ⑧ NHS economic Evaluation database

이 시기는 어떤 주제에 대해 리뷰할 지에 대해 결정하고 실제 진행할 프로토콜을 작성하는 시기이다.

작성할 주제는 PICO 질문에 맞추어 해당 주제를 정한다. PICO는 환자 혹은 문제(Patients), 중재(Intervention), 비교중재(Comparison), 결과(Outcome)에서 따온 것으로 리뷰하고자 하는 질문을 대담가능한 적절한 질문 형태로 바꾸어 주는 것을 말한다. 주제를 결정하였으면 해당 주제에 대한 기존의 체계적 고찰이 있는지 검색해 본다. 이러한 검색을 특히 Scoping search라고 한다. 그 결과 이미 체계적 고찰이 있으면 개선 가능성을 검토하고 높은 질의 체계적 고찰이 이미 있다면 대체 질문을 검토하여야 한다. 체계적 고찰을 하기로 결정하였으면 해당 주제에 대한 프로토콜을 작성한다. 프로토콜을 체계적 고찰의 배경, PICO, 연구 디자인, 방법에 대해서 5-6페이지 정도로 작성하는 것이다. 이러한 프로토콜을 작성하는 이유는 체계적 고찰의 배경을 이해하고 정제된 질문을 만들어 낼 수 있으며 계획된 검색 전략과 확실한 포함/배제 기준을 결정하고 원하는 자료를 얻는 방법을 명백히 하기 위함이다. 프로토콜을 작성하였으면 해당 주제에 대해 간단한 검색만을 실시하고 해당 주제에 대한 몇몇 논문만을 가지고 실제로 자료 추출

### III. 체계적 고찰 과정

체계적 고찰은 아래 그림과 같은 세 가지 시기로 구분할 수 있다(그림 2).

#### 1. 제 1기 :리뷰의 계획

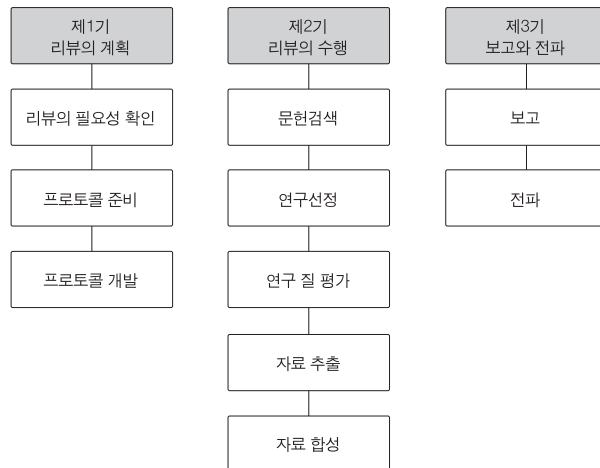


그림 2. 체계적 고찰 과정

과 자료 분석을 실시하고 그러한 경험을 바탕으로 프로토콜을 갱신하고 리뷰를 시작하면 된다.

## 2. 제 2기: 리뷰 시행

리뷰는 다음과 같은 4가지 과정을 거친다.

- (1) 자료 확인(Identification)
- (2) 논문 선정(Selection)
- (3) 결과 추출(Abstract)
- (4) 결과 분석(Analysis)

### 가) 문헌검색

문헌 검색은 검색의 범위, 검색 전략, 검색어에 대한 결정이 필요하다. 우선 검색의 범위에 대해서는 COSI라는 틀이 쓰인다.

COSI는 Core, Standard, Ideal의 약자이다. 이중 Core는 문헌 검색의 핵이 되는 부분으로 우리의 경우 관련 국내 문헌, 핵심 데이터베이스(PubMed, Embase, Cochrane central)과 포함 연구의 참고 문헌이 여기에 해당한다. Standard는 표준 검색 범위라고 할 수 있는데 핵심 잡지에 대한 수기 검색, Core에 있는 데이터 베이스 이외의 다른 일반적 데이터 베이스(Web of Science, Biosis, LILACS, DARE) 그리고 특이적 데이터 베이스(CINAHL, PsycINFO, ERIC, CANCERLIT, TOXNET, AIDSLINE)을 말한다. 체계적 고찰의 경우 최소한 standard 이상의 검색 범위를 요구한다.

Ideal은 이상적인 검색 범위로 학술대회 초록집, 출판되지 않은 문헌, 회사 접촉, 현재 진행되고 있는 임상시험 등이 포함된다.

검색 전략은 민감도 검색을 할 것인지 혹은 특이도 검색을 할 것인지이다. 민감도 검색은 해당 주제 논문이 빠짐없이 검색되도록 하는 것으로 비록 덜 적절한 논문도 포함되는 것을

감수하고서라도 논문이 빠짐없이 검색되도록 하는 전략이다. 이는 주로 체계적 총설, 임상진료 지침 개발 때 검색 전략이며, 총설, 증례보고에도 이용된다.

### 나) 자료 선정

자료를 선정할 때 첫 번째 고려 사항은 몇 명이 평가를 할 것인가이다. 여러 명의 검토자가 동시에 선정을 하면 비뚤림을 줄이고 오류를 최소화할 수 있지만 검토자간의 불일치 문제에 직면할 수 있다. 검토자가 여러명인 경우 사전에 불일치가 생길 때 이를 해결할 수 있는 과정이나 절차에 대해서 결정해 놓아야 한다. 일반적으로 주관적인 평가가 필요한 경우 최소한 두 명은 평가에 참여하여야 한다. 특정 영역의 전문가들은 미리 자신의 의견을 가지고 있어서 평가의 오류를 초래하는 경우도 있지만(7), 해당 주제에 대해 전문적인 지식이 없는 경우에 비해서 연구의 타당성 평가에 있어서 더 일관성이 있다(8). 연구의 질을 평가할 때 저자, 소속, 잡지, 연구 결과 등에 대해 맹검을 적용할지를 결정하여야 한다. 일부 연구에 의하면 맹검을 하면 점수가 낮아지며 일관성이 증가하였지만(8) 이득이 없다는 연구도 있다(9).

### 다) 질 평가

개별 문헌에 대한 질평가는 문헌 고찰 과정에서 비뚤림을 최소화하는데 그리고 결과를 해석하는데 필요하다. 평가는 대체로 결과의 적용성, 연구의 타당성 그리고 결과 해석에 영향을 미치는 연구 디자인 등이 이루어진다. 이중 연구의 타당도가 대체적으로 연구의 질에 해당한다. 연구의 타당도란 비뚤림을 최소화 하기 위해 연구 기획이나 실행에서 시행한 노력의 정도를 말하며(10), 타당도의 변이는 결과의 변이를 일부 설명할 수 있다. 적절한 연구논문이 선정되면 각 연구결과로부터 적절한 자료를 추출해야 한다. 임상 시험에서 비뚤림의 원천은 선택 비뚤림(selection bias), 실행 비뚤림(performance

bias), 탈락 비뚤림(attrition bias), 결과 확인 비뚤림(Detection bias)의 네 가지이다. 질 평가 기준이 실제 연구 결과의 질을 반영한다는 명백한 근거는 부족하지만 논리적인 근거는 명백하며 그런 이유로 문헌 평가에 쓰일 수 있다(11).

라) 분석

분석은 결과는 연구 결과들이 비교적 일관성이 있고 동질성이 있는지 여부에 따라 달라진다. 만일 연구들이 동질성이 있으면 연구를 합해서 전체 효과를 산출하면 된다. 하지만 이질성이 있을 때는 이질성이 생기는 이유에 대해 탐구하고 이에 대한 적절한 해결을 하여야 한다. 이질성은 각 연구들 간의 변이성 정도가 어떠한가를 평가하는 것이다.

만약 통합된 연구결과가 매우 다양하여 이질적인 결과를 보인다면 이는 연구들이 서로 동질하지 않을 가능성이 있는 것이며, 메타분석에 적합하지 않다. 이질성을 해결하는 방법에는 하부집단 분석, meta-regression 등이 있으며 random effect model을 이용해서 보수적인 결론을 내리기도 한다. 고정 효과모형(Fixed effects models)은 한 연구내의 변이성만 고려하는 모형인데 비해서 무작위 효과모형(Random effects models)은 연구 내 변이성뿐만 아니라 연구간 변이성도 고려

하는 모형이다. 어떻게 자료에 대한 값을 구할지는 자료를 합칠 수 있을 지와 자료의 형태와 이산변수인지 연속 변수인지 연속 변수라면 같은 측정도구를 사용하였는지에 따라서 달라진다(12)(그림 3).

필요에 따라서 민감도 분석을 실시할 수 있다. 민감도 분석이란 어떤 특정 연구만, 특정 환자만, 또는 특정 치료만 따로 분리하여 분석하는 것이다. 포함기준, 연구의 질 등에 따라서 분석을 하여도 비슷한 결과가 나온다면 메타 분석의 신뢰도를 높일 것이다.

체계적 고찰을 통해 비뚤림 없는 결론에 도달하기 위해서는 해당 연구 주제에 관한 대부분의 일차 연구들이 포함되어야 한다. 하지만 시행된 모든 연구들이 출판되지 않기 때문에 문제가 발생한다. 연구 중 유의한 결과를 보이는 것들이 더 많이 투고되고 더 많이 출판되며 더 빨리 출판되는 문제가 있다. 따라서 출판된 것만 가지고 분석하면 결과가 과장되게 되는데 이를 출판 비뚤림이라고 한다. 출판 비뚤림을 확인하는 방법으로 가장 많이 이용되는 것은 Funnel plot이다. Funnel plot은 effect size와 sample size에 대한 plot으로 세로축에 1/SE를 사용하기도 한다. 출판 비뚤림에 대한 통계 검정으로는 Rank correlation test, Linear regression test 등을 이용한다(6).

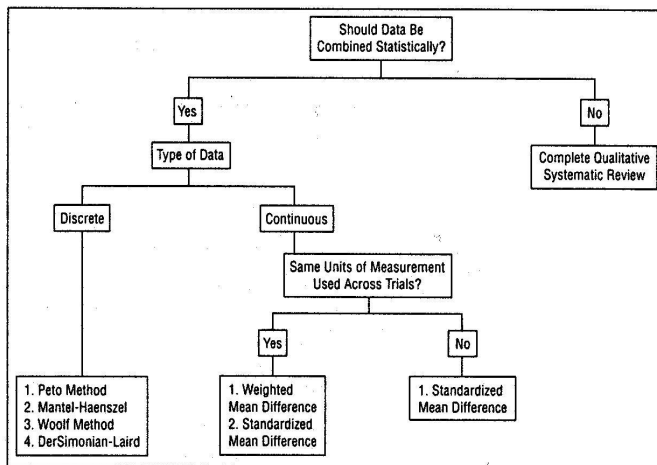


그림 3. 체계적 고찰에서 통계값 선택의 알고리즘

## IV. 요약

체계적 고찰은 명확히 정의된 문헌검색에 기반하고, 결정된 기준(criteria)을 이용하여 선택된 문헌의 질을 평가하며, 타당성이 입증된 방법으로 분석을 시행하는 리뷰의 한 형태이다. 체계적 고찰은 (1)대답을 얻고자 하는 질문을 공식화 하고 이에 대한 프로토콜을 작성하고, (2)해당 질문에 대한 일차 연구를 찾아서 포함시키고 (3)자료의 질을 평가하고, (4)자료를 추출하여 뽑고, (5)자료를 분석하며, (6)결과를 해석하고 보고서를 작성하는 순서로 이루어진다 하지만 체계적 고찰 혹은 메타분석이 모든 문제에 대한 해결책은 아니다. 메타분석이 수행되었다고 하더라도 질이 나쁜 연구가 포함되거나 질 문제를 무시한 경우, 이질성에 대해 적절한 주의를 기울이지 못할 때, 무차별적인 자료 병합으로 결론이 정확하지 않은 경우, 출판 비뚤림, 시간 지체 비뚤림, 이중 출판 비뚤림, 언어 비뚤림, 결과 보고 비뚤림과 같은 보고 비뚤림이 문제가 되는 경우에는 메타분석 혹은 체계적 고찰이 잘못된 결론으로 도달하도록 할 수 있다. 따라서 올바른 방법론에 기초한 체계적 고찰만이 현존하는 모든 근거를 결합하여 올바른 결론으로 이끌 수 있다.

### 참고문헌

1. 한림의대 가정의학교실편. 근거중심의학의 이론과 실제, 제 1판. 서울; 고려의학; 2002.
2. Klassen TP, Jadad AR, Moher D. Guides for reading and interpreting systematic reviews: I. Getting started. Arch Pediatr Adolesc Med. 1998; 152(7):700-704.
3. Agency for Healthcare Research and Quality. Systems to Rate the Strength Of Scientific Evidence. Evidence Report/Technology Assessment Number 47 2002.
4. Egger M, Smith GD, Sterne JA. Uses and abuses of meta-analysis. Clin Med 2001; 1(6):478-484.
5. Counsell C. Formulating questions and locating primary studies for inclusion in systematic reviews. Ann Intern Med 1997(5); 127:380-387.
6. The Cochrane Collaboration. Cochrane Handbook for Systematic Reviews of Interventions 4.2.5. Updated May 2005
7. Oxman AD, Guyatt GH. The science of reviewing research. Ann NY Acad Sci 1993; 703:125-133.
8. Jadad AR, Moore RA, Carroll D, et al. Assessing the quality of reports of randomized clinical trials: Is blinding necessary? Controll Clin Trials 1996; 17(1):1-12
9. Berlin JA. University of Pennsylvania meta-analysis blinding study group. Does blinding of readers affect the results of meta-analyses? Lancet 1997; 350(9072):185-186.
10. Moher D, Jadad A, Nichol G, Penman M, Tugwell T, Walsh S. Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists. Controll Clin Trials 1995; 16(1):62-73.
11. Feinstein AR. Clinical Epidemiology: The Architecture of Clinical Research. Philadelphia: Saunders, 1985: 39-52.
12. Moher D, Jadad AR, Klassen TP. Guides for reading and interpreting systematic reviews: III. How did the authors synthesize the data and make their conclusions? Arch Pediatr Adolesc Med. 1998; 152(9):915-920.