대한내분비학회지: 제 23 권 제 6 호 2008
10.3803/jkes.2008.23.6.361

메타분석

고려대학교 의과대학 의학통계학교실

이 준 영

Meta-analysis

Juneyoung Lee

Department of Biostatistics, College of Medicine, Korea University

The Rosiglitazone Story-계속되는 불확실성의 세계

2007년 6월에 Nissen과 Wolski는 제2형 당뇨병 환자들에게 사용하는 혈당강하제 중 하나인 rosiglitazone (Avandia, GlaxoSmithKline)의 안정성 여부를 평가한 메타분석 연구결과를 NEJM에 발표하였다[1]. 주 평가변수는 심근경색(myocardial infarction, MI) 발생률 및 심혈관 질환으로 인한 사망률(cardiovascular death, CVD)이었으며, 메타분석에 사용한 연구들은 rosiglitazone 투여군과 위약군 또는 활성 대조군 간의 심혈관계 질환에 대한 이환 및 사망 건수를 보고한 42개의 무작위 임상시험(randomized clinical trial, RCT) 연구들로, 이 중 40개는 소규모 임상시험 연구들이었으며 2개는 대규모 임상시험인 The Diabetes Reduction Assessment with Ramipiril and Rosiglitazone Medication (DREAM) trial 및 A Diabetes Outcome Prevention Trial (ADOPT)이었다. Nissen과 Wolski는 소위 Peto 방법을 사용한 메타분석을 통해 rosiglitazone군의 MI 발생 odds는 위약군 또는 활성 대조군에 비해 1.43배(95% 신뢰구간, 1.03~1.98) 그리고 CVD 사망 odds는 1.64배 (0.98~2.74) 높다고 발표했다. 이를 근거로 Psaty와 Furberg은 이제 rosiglitazone의 안전성 여부는 불분명하며, 따라서 FDA가 이후 개발되는 제2형 당뇨병 치료제 신약을 허가할 때에는 대리 종료점(surrogate end point)에 기초한 단기 연구결과 이외에도 추가적인 대규모, 장기 제4상 임상시험 연구결과를 조건부로 허가하는 것을 고려할 필요가 있고[2], 필요하다면 이에 대한 의무조항 신설을 고려해야 할 것이라고까지 주장하였다[3]. Nissen과 Wolski의 논문으로 인해 심혈관계 질환으로 인한 입원율 및 사망률에 대해 rosiglitazone과 메트포르민(metformin) 또는 설폰요소제 (sulfonylurea)의 병합요법을 메트포르민과 설폰요소제 병합요법과 비교하는 The Rosiglitazone Evaluated for Cardiac Outcome and Regulation of Glycaemia in Diabetes (RECORD) trial이 원래 계획에도 없던 중간분석을 실시하게 되었고, 그 결과 rosiglitazone군에서 심부전 발생 위험이 더 높게 나타나긴 했지만, 두 군 간의 심혈관계 질환 입원율 및 사망률 차이는, 중간분석으로 인한 낮은 검정력과 예상보다 높게 나온 추적실패율 및 낮은 발생률로 인해, 아직은 확증적인 결론을 내릴 수 없으며, 자체 Data and Safety Monitoring Board (DSMB)에서는 이 연구를 계속적으로 진행하도록 추천했다고 발표하였다[4]. 하지만 이는 곧 RECORD trial 결과가 Nissen과 Wolski의 메타분석 결과와는 상이하지 않음을 의미하는 것이라는 의견[5] 및 혈당 조절이 미세혈관 합병증 억제에는 효과적이긴 하지만 이와는 별개로 제2형 당뇨병 치료 시 굳이 대혈관 부작용을 유발할 수 있는 약물을 사용할 필요가 있겠는가라는 주장[6]도 제기되었다. FDA advisory committee meeting에 참여했던 Rosen (2007)은 이제 당뇨병 치료제에 대한 평가를 위해서는 혈당 조절 대리 종료점인 HbA_{1c}가 아닌 장기간에 의해 관찰되는 임상적 종료점을 사용해야 할 것이라고 언급했다[7]. 하지만 이 주장에 대해 만일 그렇게 한다면 당뇨병 치료제와 관련된 신약개발이 심각하게 지체되게 될 것이기 때문에 FDA에서는, 심혈관계 관련 안전성에 관한 문제 제기가 되는 경우에 한해서만, 시판 전 리뷰와 시판 후 조사를 통한 안전성 여부 평가를 고려하고 있다고 정정했고(Joffe HV 등), GSK에서는 FDA 자체적으로 실시한 메타분석 결과, rosiglitazone의 MI 발생 위험이 위약군과 비교했을 때에만 통계적으로 유의했지 메트포르민이나 설폰요소제 투여군과 비교했을 때에는 차이가 없었을 뿐더러, GSK 자체의 역학조사 결과도 rosiglitazone은 pioglitazone과 비교해 MI 발생 위험에 유의한 차이가 없었다고 반박하였으며(Krall RL); 이에 대해 Rosen은 FDA와 관련된 언급은 개인적인 의견이었을 뿐이며, GSK의 역학조사 결과는 그 자체적으로

한계가 있는 연구결과라고 답변하였다(Rosen CJ)[8]. 한편 메타분석 시 Nissen과 Wolski가 사용했던 Peto 방법은 다른 여러 메타분석 방법 들 중 rosiglitazone의 MI 발생 위험에 관해 유일하게 통계적으로 유의한 결과를 제공해 주는 방법인데, 이 메타분석 자료처럼 비교집단 간 환자 수가 불균형을 이루는 연구결과들을 결합하고자 하는 경우에는 Peto 방법을 사용하는 것이 부적절하다(Bracken MB). 메타분석에서 사용된 DREAM trial의 경우 당뇨병이 발병하는 환자들에 대해서는 더 이상의 관찰을 중지했었고, ADOPT trial은 혈당 조절 수준이 만족스럽지 않은 경우에 대해서는 추가 관찰을 중지했었기 때문에 두 trial 모두 주 평가변수에 심혈관계 사건 발생여부가 포함된 연구들은 아니며, 또한 두 trial 모두 rosiglitazone 군의 추적기간이 대조군에 비해 상대적으로 더 긴 데에도 불구하고 메타분석 시에는 이를 감안하지 않았기 때문에 해당 메타분석 결과는 과대 추정되었을 가능성이 있다(Mannucci E 등). 두 군 간 절대위험의 차이는 0.2% 정도밖에 되지 않는데, 분석에 사용된 42개 무작위대조군시험(RCT)들 중 MI 발생률에 대해서는 event가 발생하지 않았던 4개 trial이, CVD 사망률에 대해서는 19개 trial이 분석에서 제외되었기 때문에 이들 결측값들이 결과에 영향을 주었을 수 있다. 따라서 전체 자료를 모두 사용하는 베이지안 메타분석을 실시해 본 결과, 두 경우 모두 통계적인 차이는 없는 것으로 나타났다(Diamond GA, Kaul S). 이런 주장과 함께 임상 가이드라인 및 이러한 진료기준을 마련하는 전문가들 모두 제2형 당뇨병 치료의 주 목적으로 혈당조절을 제시하고 있기 때문에 임상의들이 rosiglitazone을 처방하는 것은 잘못 되었다는 Psaty와 Furberg (2007)의 주장이야 말로 잘못된 것이다(Brett, A) 등 rosiglitazone 메타분석 결과에 대한 반론들이 제기되었으며, 이에 대해 원 저자들은 Peto 방법은 event 수가 적은 연구들을 결합하는 경우에 가장 적절한 메타분석 방법인 것으로 이미 알려져 있으며, DREAM trial과 ADOPT trial 모두 비교집단들 간 노출기간에 차이가 있다는 증거는 없고, 해당 안전성 자료들은 무작위화가 실시된 ITT 분석대상 자료들이기 때문에 이미 두 집단 간에는(예후요인에 관한) 균형이 이루어졌었을 뿐만 아니라, 베이지안 메타분석은 여전히 논란의 여지가 있는 통계적인 방법이고, 이 방법에 의하면 언제나 신뢰구간이 넓게 나올 수 밖에 없다고 반박하였다[9]. 최근 들어서는 혈당조절 강화요법은 표준요법보다 사망률이 높았으며 두 요법 간에 주요 심혈관계 사건 발생률의 차이는 없었다는(그리고 rosiglitazone과 관련해 유의한 심장 부작용 효과는 관찰되지 않았다) The Action to Control Cardiovascular Risk in Diabetes (ACCORD) trial[10] 및 혈당조절 강화요법이 표준요법에 비해 주요 대심혈관계 및 미세혈관계 사건 발생율을 10%까지 낮추었다는 The Action in Diabetes and Vascular Disease: Preterax and Diamicron Modified Release Controlled Evaluation (ADVANCE) trial[11] 결과가 NEJM에 동시에 발표되는 등 당뇨병 치료제의 심혈관계 안정성에 관한 논란은 2008년 현재에도 여전히 지속되고 있다[12].

이러한 모든 혼돈(?)의 시발점이 된 연구가 메타분석이라는 통계적 방법을 사용한 연구라는 점에서 이 방법에 대한 관심이 증가되는 것은 당연하다고 할 수 있겠다. 하지만 메타분석 연구의 주체는 통계 전문가가 아니라 임상연구자이기 때문에 메타분석 방법론의 장단점을 충분히 파악하고 분석 결과를 정확히 해석하기 위해서는 먼저 이 연구방법론에 대한 전반적인 이해가 필요하다고 본다.

서 론

Meta-analysis refers to the analysis of analyses. I use it to refer to the statistical analysis of a large collection of results from individual studies for the purpose of integrating the findings. It connotes a rigorous alternative to the causal, narrative discussion of research studies which typify an attempt to make sense of the rapidly expanding research literature.

Glass (1976)

의학 연구논문에 제시된 자료들의 결합이 처음으로 시도된 것은 1904년 Karl Pearson에 의해서이며[13,14], 공식적인 메타분석 결과가 처음으로 발표된 연구는 1954년 Lasagna 등의 연구이다[14,15]. 메타분석이라는 용어는 1976년 Gene Glass가 통계학적 방법이 아닌 철학적 개념으로 이 용어를 사용하면서 처음으로 언급하였다[16]. Glass는 문헌들을 고찰(review)하는 것 그 자체가 하나의 연구방법론에 해당하며, 따라서 이 고찰 역시 다른 공식적인 연구에 못지않게 체계적으로(systematically) 수행되어야 한다고 주장했다. 한편 의료행위 중 많은 부분이 적절한 근거를 기반으로 이루어지지 않고 있음을 설파한 Archie Cochrane은 메타분석 분야의 선구자 중 한 사람이라 할 수 있다. 이 영국의 역학자는 1972년 그의 유명한 원고 Effectiveness and efficiency에서 의료와 관련된 의사결정을 하고자 하는 많은 사람들(policy makers, professionals, consumers of services)이 현재 가용하고, 믿을만하며, 엄격하게 정리된 근거들을 쉽게 접근할 수 있는 방법을 제공받지 못하고 있다고 언급하였다[17,18]:

It is surely a great criticism of our profession that we have not organised a critical summary, by speciality or subspeciality, adapted periodically, of all relevant randomized controlled trials.

*Cochra*ne (1979)

이러한 Cochrane의 비판은 출산과 신생아에 대한 적절한 관리방법을 연구한 RCT 결과들을 정리하는 10년간에 걸친 국가 간 협력연구를 점화시켰고, 결과적으로 모든 보건 및 의료 관련 연구결과들을 데이터베이스화하고자 하는 임상연구자들, 연구방법론 전문가들 및 사용자들 간 국제적 네트워크인 코크란 협력단체(Cochrane Collaboration)를 탄생시키는 밑거름이 되었을 뿐만 아니라[19], 궁극적으로는 근거-기반 의학(evidence-based medicine, EBM) 운동을 점화시키는 도화선 역할을 하게 되었다[20]. EBM은 "각 환자들에게 필요한 치료방법을 결정할 때, 현존하는, 가장 최신의, 그리고 가장 최상의 근거를 진지하고, 명료하며, 사려 깊게 사용하는 것"이라 할 수 있다[21]. 이를 위해 두 종류의 근거, 즉, 임상가 본인의 전문적 지식인 내적 근거와 다른 연구자들의 연구결과인 외적 근거들을 종합해 임상적인 의사결정을 하고자 하는 것이다. 따라서 EBM을 시행하기 위해서는 환자들을 보살피는 데 도움이 되는 연구들을 찾아내고, 이 연구결과들을 제대로 평가할 수 있어야 할 것이다[22]. 물론 RCT가 EBM에 있어서 최상의 기준이긴 하지만, 체계적 고찰(systematic review) 역시 EBM에 효과적으로 사용될 수 있는 방법이다.

체계적 고찰은 명백한 연구목적 하에 과학적인 방법으로 수행된 일차 연구들을 대상으로, 연구주제 설정, 연구들의 포함기준, 자료수집 방법 등을 미리 계획하고 사전 정의한 다음, 이를 객관적이고 재현성이 확보된 방법론을 사용해 찾아낸 후, 이들의 연구결과에 대한 고찰과 분석(즉, 결과의 합성)을 실시하여, 연구질문에 대한 결론을 내리고 신뢰성 있는 정보를 제공하고자 하는 연구합성 방법론(research synthesis methodology)이다[23]. 체계적 고찰에서는 통계적인 방법을 사용해 각 연구의 상대적인 공헌도를 감안하면서 자료들을 결합할 수도 있지만, 연구들 간의 특성이 너무 이질적이면 자료의 결합을 시도하지 않을 수도 있다. 이러한 자료결합 단계를 메타분석이라 한다. 체계적 고찰이라는 용어와 메타분석이라는 용어가 서로 혼용되어 사용되기도 하지만, 엄밀히 말한다면 메타분석은 체계적 고찰의 틀 안에서 수행되는 연구결과 결합방법론이라 할 수 있다. 메타분석이라는 용어는 글자 뜻 그대로 - meta는 그리스어로 'after', 'beyond', 'with' 라는 뜻을 나타내는 전치사 - 하나의 주제에 대해 연구한 여러 연구의 분석결과들을 분석하는 것(analysis of analyses)을 의미하며, 따라서 메타분석은 특정 중재와 관련해 기존에 출판되었거나 아직 출판되지 않은 모든 연구결과들을 융합하고자 하는 통계적인 연구방법론을 의미한다. 메타분석은 대규모 연구들뿐만 아니라, 사용된 연구방법, 연구대상자들의 특성, 결과에 영향을 미칠 수 있는 다른 중요한 요인들 등에 있어서 변동을 보이는, 소규모로 진행된 연구들에 대해서도 사용될 수 있는 방법론이다. 그렇다면 이들 소규모 연구들을 결합한 메타분석 결과와 의학적인 중재효과를 증명할 수 있는 최상의 기준으로 간주되는 대규모 RCT 결과가 서로 다르다면 이를 어떻게 받아들여야 하는가? 어쩌면 이보다 더 중요한 점은, 이들 중 하나만 있는 경우에, 해당 연구의 결과인 신뢰구간(결과에 대한 신뢰의 정도)이나 표준오차(결과에 대한 불확실성의 정도)를 얼마나 신뢰할 수 있겠는가 하는 점일 것이다. 따라서 이에 대한 결론은 각 연구의 질과 분석의 정확성에 달려있다고 할 수 있다.

메타분석의 목적은 활용 가능한 연구들의 장점 및 제한점들에 기초하면서 객관적인 결론을 유도하고자 하는 것이다[14]. 즉, 유사한 결과를 보기 위해 유사한 환자들을 대상으로 유사한 처리를 사용한 유사한 연구들의 자료를 결합함으로써 가능한 한 최대의 통계적 검정력(power)과 정밀성(precision)을 확보한 상태에서 해당 처리효과를 추정하고자 하는 것이다[24]. 메타분석에는 다수의 임상시험 연구모집단(study population)들이 포함되기 때문에 표본수가 증가하게 되고, 따라서 만일 실제로 처리 간에 차이가 존재한다면 이를 더 쉽게 탐지할 수 있게 되며(즉, 검정력이 증가하며) 또한 처리 간 차이를 추정하는 데 있어서도 정밀성이 증가하게 된다(즉, 신뢰구간이 좁아진다). 물론 메타분석을 잘 수행하기 위해서는 시간도 많이 소요되고 비용도 많이 들긴 하지만, 그래도 대규모 임상시험에 비하면 훨씬 적게 드는 것이 사실이다.

메타분석에서는 각 연구들이 원래 추구했던 임상질문들에 대한 답을 하기 위해 두 개 또는 그 이상의 독립적인 연구결과(주로 처리 간 평균 차이 및 해당 신뢰구간)들을 특정 방법, 즉, 가중평균방법을 사용해 결합한다. 이때 가중치로는 각 연구들의 표본 크기(sample size), 정밀성(precision, 예를 들면 처리 간 평균 차이에 대한 표준오차), 또는 각 연구의 질(study quality) 등이 주로 사용된다. 하지만 메타분석의 핵심은 해당 주제와 관련된 정보들을 얼마나 최대한 그리고 얼마나 체계적으로 찾아내서 요약했는가에 있다. 제대로 된 메타분석을 수행하기 위해서는 연구주제에 대한 전문적인 지식뿐만 아니라 해당 방법론에 대한 전문적인 지식 또한 필요하다. 따라서 메타분석은 임상전문가와 경험이 있는 통계전문가 간의 공동연구로 진행되는 경우가 거의 대부분이다. 연구주제와 관련된 선정기준에 해당하는 연구들을 최대한으로 찾아내고 메타분석 과정 중에 발생하게 될 불확실성을 줄이기 위해서는 무엇보다도 먼저 메타분석 주제에 관한 명확한 설정이 필요하다. 그 다음으로 선정기준에 부합하는 모든 적절한 연구결과들을 파악하기 위한 문헌탐색이 광범위하게 이루어져야 한다. 보통의 경우 이렇게 탐색된 문헌들의 수가 매우 많기 때문에 이들이 메타분석 포함기준에 적절한지 평가하기 위해 먼저 초록 검토(abstract review)를 실시한 후, 그 다음으로 자세한 평가와 자료추출 그리고 연구의 질 평가를 위해 전체를 검토(full review)하는 과정을

거치게 된다. 일반적으로는 이러한 과정을 거치면서 대략 탐색된 문헌들의 10% 전후만이 남게 된다. 자료추출 및 질 평가를 위한 full review 과정은 미리 준비된 양식지를 사용해 조심스럽고 세밀하게 진행되어야 한다. 마지막으로 추출된 자료의 결합을 시도하게 된다. 물론 연구들 간의 특성이 너무 차이가 나서 이러한 자료결합이 무의미한 경우에는 최종 선정된 각 개별 연구결과들을 요약 정리하는 것으로 메타분석을 마무리할 수도 있지만, 대부분의 경우에는 이 자료들을 적절한 통계적 방법을 사용해 하나의, 대표적인, "최선의" 추정치 및 신뢰구간으로 요약하게 된다. 사실 이 마지막 단계가 메타분석이 비판을 받는 주된 부분이기도 하다. 왜냐하면 메타분석에 포함된 연구들이 실질적으로는 서로 다른 것을 측정하고 있다는(즉, 연구결과들이 이질적이라는) 임상적, 통계적 증거들이 있는 경우가 많기 때문에 자료결합이 의미가 있기 위해서는 추가적인, 그러나 논란의 여지가 있는, 통계적 가정을 해야만 하기 때문이다(이에 대해서는 5) 통계분석 및 결과 제시에 있는 (2) 변량효과모형을 참조).

1. 메타분석 관련 프로그램

메타분석을 수행하기 위한 프로그램들은 많이 있다. 그 중 가장 많이 알려진 것들로는 코크란 협력단체에서 무료로 제공하는 RevMan (Review Manager, Version 5.0, http://www.cc-ims.net/RevMan) 프로그램, 유료이면서 메타분석 전문 프로그램인 Comprehensive Meta-Analysis Software (Version 5.0, http://www.meta-analysis.com), 그리고 상용 통계분석 전문 소프트웨어이면서 고급화된 메타분석 수행이 가능한 STATA (Version 10.0, http://www.stata.com) 등을 들 수 있다. 본 고찰에서는 STATA 프로그램을 사용해 메타분석의 예를 설명해 나가기로 하겠다.

2. 메타분석의 수행

메타분석은 보통 다음과 같은 다섯 단계에 걸쳐 수행하게 된다[25,26].

① 연구주제 설정(formulating research questions)
② 연구 포함기준 및 제외기준 설정(defining inclusion and exclusion criteria)
③ 문헌탐색 및 논문선택(literature search and study selection)
④ 연구의 질 평가 및 자료추출(study quality assessment and data extraction)
⑤ 통계분석 및 결과 제시(statistical analysis and result presentation)

1) 연구주제 설정

메타분석의 첫 단계는 연구주제를 명확히 설정하고 연구목적과 가설들을 자세히 명시한 프로토콜을 작성하는 것이다. 연구주제는 당연히 임상적으로 관심이 있는 것이어야 하겠지만, 주로 확증된 것 보다는 논란이 많은 주제가 선택되게 된다. 연구의 목적은 구체적이고 명확해야 되며 중재의 종류, 비교집단에 관한 내용, 임상적 관심변수 등이 포함되어야 한다[27]. 예를 들면, "제2형 당뇨병 환자에 대한 rosiglitazone의 단독 혹은 병합요법이 위약 및 인슐린, 설폰요소제, 메트포르민 등 다른 혈당강하제의 투여와 비교해 심혈관계 질환의 이환율 및 사망률을 높이는 결과를 가져오는가?" 등이다. 코크란 가이드라인에서는 이와 같이 임상적 문제를 구조적인 질문의 형태로 전환하여 연구의 개념적 틀을 구축하기 위해 다음의 PICO 기준에 기초한 핵심질문 설정을 추천하고 있다[28]: **P**atient population (어떤 집단에 대해 연구할 것인가?) / **I**ntervention 또는 exposure (어떤 중재에 대한 연구를 하려고 하는가?) / **C**omparison (어떤 비교대상을 사용해 해당 중재가 더 좋은 결과를 제공한다는 것을 보이고자 하는가?) / **O**utcomes (해당 중재의 효과로 무엇을 보고자 하는가?).

이와 더불어 혹시 해당 주제와 관련된 연구의 수가 너무 많아 메타분석을 진행하기 힘들 정도는 아닌지, 그리고 연구간에 임상적, 방법론 적 이질성(heterogeneity)이 너무 심해 이 연구결과들을 결합하는 것이 혹시 무의미한 작업은 아닌지 등에 대해서도 이 단계에서 미리 심사숙고해야 한다.

2) 연구 포함기준 및 제외기준 설정

연구주제에 합당한 모든 적절한 연구들을 찾아내기 위해서는 먼저 연구 선정기준을 자세히 정의해야 한다. 여기에는 주로 연구설계의 형태(예: RCT로 한정할 것인지 아니면 코호트 연구 등 관찰연구도 포함할 것인지), 환자군의 특성(예: 연령, 성별, 의학적 상태 여부 등), 자료의 출판형태(예: 초록집에만 출판된 자료를 포함시킬 것인지, 미출판 자료 역시 포함시킬 것인지), 사용된 언어(예: 영어권 이외에 출판된 연구들을 포함시킬 것인지), 그리고 연구기간(예: 어느 기간 동안에 수행된 연구들로 한정할 것인지) 등이 포함된다. 일반적으로 메타분석 대상은 무작위 이중눈가림 임상시험들이다. 왜냐하면 무작위 할당(random allocation)이 이루어져야만 bias가 최소화되거나 집단 간에 bias의 균형이 이루어져서 예후에 대한 비교성이 확보될 수 있기 때문이다[21].

포함기준은 주 연구주제에 관한 결론을 내릴 수 있게끔 구성되어야만 한다[26]. 예를 들어, 삶의 질에 관한 처리효과를 보고자 한다면 결과(outcome)로 삶의 질을 평가하지 않은 연구들은 제외되어야 할 것이다. 사실 포함기준을 어떻게 설정하는지가 메타분석 결과에 막대한 영향을 끼치게 된다. 이를 지나치게 넓게(진보적으로) 설정하게 되면(예를 들어 비 무작위 임상시험이나 관찰연구 등을 포함시키면) 보다 광범위한 연구를 수행할 수 있긴 하지만 좀더 주관적인 결론을 내리게 될 가능성이 있고, 연구들 간의 이질성 또한

커져서 메타분석 결과의 신뢰성이 낮아지게 될 것이다. 반면에 이를 너무 좁게(보수적으로) 설정하면 해당 연구는 제한된 형태의 부집단 분석연구가 될 가능성이 있다[29].

3) 문헌탐색 및 논문선택

연구주제에 관한 답을 하기 위해서는 주제와 관련이 있는 연구들을 가능한 한 많이 찾아내는 것이 중요하며, 설정된 포함기준에 해당하는 연구들을 모두 찾아내도록 노력해야 한다. 또한 문헌탐색 시 사용된 특정 탐색어들은 연구결과 보고서(혹은 논문)에 제시해야 한다. 효과적인 문헌탐색을 위해서는 아래의 전략들을 단계적으로 사용하는 것이 좋다.

① Cochrane Library, MEDLINE, EMBASE, BIOS (Life sciences and Biomedicine), SCI (Science Citation Index), CINAHL (Nursing and Allied Health Literature), AMED (Allied and Alternative Medicine), Biosis (Biological Abstracts), PsycINFO (Psychological Abstracts) 등 전자 의료관련 데이터베이스를 사용해 상호심사과정을 거친 논문(peer-reviewed journal)을 찾는다.

② 해당 주제와 관련된 출판서적이나 관련 학술회의의 초록집 등을 검토한다.

③ 국내외 관련 학회의 웹사이트 등 특정 인터넷 사이트를 조사하고, 해당 질병과 관련된 진료 가이드라인 등을 살펴본다.

④ 전국 혹은 지역 단위의 등록체계(registry)를 조사하고 미출판된 논문들을 찾는다. 이를 위해 해당 질병과 관련해 명망 있는 학자나 의료센터 등을 접근해 본다.

⑤ 출판되었거나 미출판된 연구자료에 대한 정보를 입수하기 위해 제약회사의 의료관련 담당자들과 연락을 취한다.

⑥ Intute (www.intute.ac.uk) 등과 같은 의료관련 포탈사이트 및 Google, Yahoo 등과 같은 일반 포탈사이트를 조사한다.

전자 의료데이터베이스 탐색 시에는 포함기준을 만족시키는 출판논문들을 빠트리지 않기 위해 적어도 두 개 이상의 데이터베이스를 동시에 탐색하는 것이 매우 중요하다. 영어 이외의 다른 언어로 쓴 논문들 역시 반드시 찾아야 하며(뒤에 기술한 'Publication bias' 참조), 필요한 경우에는 번역도 실시해야 한다[30]. 초록집은 대부분 상호심사과정을 거치지 않은 연구이며 해당 연구의 질을 평가하기에 충분한 정보들도 없는 경우가 많기 때문에(뒤에 기술한 '연구의 질 평가 및 자료추출' 참조), 메타분석 결과의 민감도를 평가하는 정도로만 사용하기도 한다(뒤에 기술한 '민감도 분석' 참조). 또한 신경 써야 할 부분은 중복게재된 논문이나 이전에 발표된 자료를 확대 혹은 축소 재생산한 논문들이다. 예를 들어 하나의 RCT는 동일한 환자모집단을 대상으로 실시한 연구이지만 여러 개의 논문으로 출판되기도 하므로 이런 논문들이 중복 계수되지 않도록 주의해야 하고, 분석 시 제외해야 한다. 또한 출판되지 않았거나 출판을 준비하고 있는 부정적(negative) 연구들(통계적으로 유의한 처리 간 효과 차이를 보이지 못한 연구들)을 찾아내도록 노력해야만 한다. 이를 위해 연구관련 등록체계를 살펴보거나 해당 분야에서 명망이 있는 연구자에게 직접 연락을 취해보는 것도 효과적인 방법이다.

참고로 문헌탐색 시 RCT 관련 논문들을 찾기 위해 사용될 수 있는 탐색전략을 예로 들면 Table 1과 같다[28].

연구 선정기준에 따라 탐색된 문헌들 중에서 메타분석을 실시하기 위한 논문들을 선택할 때에는 유사한 디자인이 사용된 논문들을 선택해야 한다. 논문들 간 유사성을 평가하는데 사용될 수 있는 기준들은 다음과 같다(물론 이들 중 연구 선정기준에서 이미 결정된 사항들에 대해서는 재평가할 필요가 없다).

⑦ 연구설계의 종류{예: 평행설계(parallel design)인지 교차설계(crossover design)인지}

⑧ 각 연구의 환자모집단 포함기준(예: 관상동맥질환을 가진 심장병 환자군 대 관상동맥질환을 가지고 있지 않은 환자군, 모든 제2형 당뇨병 환자군 대 최소한 하나 이상의 심혈관계 질환 위험인자를 가지고 있는 제2형 당뇨병 환자군)

⑨ 각 연구의 환자모집단 제외기준(예: 65세 이상인 심장병 환자 또는 심장병 환자이면서 신장 질환을 가지고 있는 환자, 제2형 당뇨병 환자이면서 장기 인슐린 치료 중인 환자)

⑩ 각 연구가 평가한 중재방법(예: 베타차단제 대 위약인지 아니면 베타차단제 대 이뇨제인지, 혈당조절을 위한 단독요법 대 위약인지, 단독요법 대 병합요법인지 아니면 강화혈당조절요법 대 표준혈당조절요법인지)

⑪ 연구에 사용된 일차평가변수(예: 사망인지 재입원인지, 주요 심혈관계 질환의 발생률인지 심혈관계 질환관련 사망률인지)

Rosiglitazone 안전성(safety) 메타분석의 경우, 42개의 논문들을 찾아내기 위해 FDA 홈페이지, GSK 홈페이지 및 저널에 출판된 논문들을 조사하였다. 선택된 42개 RCT 중 5개는 해당 약제의 허가를 받기 위해 FDA advisory board hearing에 제출된 소규모 연구들이고(상호심사과정을 거치고 저널에 발표된 연구 4개 및 미발표된 연구 1개) 35개는 GSK 홈페이지의 임상시험 등록체계(registry)에 등록된 소규모 연구들(저널에 발표된 연구 9개, 미발표 연구 26개)이며, 나머지 2개는 최근에 발표된 대규모 임상시험 연구결과(DREAM trial 및 ADOPT trial)이다[1].

Table 1. Cochrane highly sensitive search strategy for identifying randomized trials in MEDLINE. Following examples are adapted from Higgins and Green (2008)

Ovid format		PubMed format	
#1	randomized controlled trial.pt.	#1	randomized controlled trial [pt]
#2	controlled clinical trial.pt.	#2	controlled clinical trial [pt]
#3	randomized.ab.	#3	randomized [tiab]
#4	placebo.ab.	#4	placebo [tiab]
#5	clinical trials as topic.sh.	#5	drug therapy [sh]
#6	randomly.ab.	#6	randomly [tiab]
#7	trial.ti.	#7	trial [tiab]
#8	1 or 2 or 3 or 4 or 5 or 6 or 7	#8	groups [tiab]
#9	animals.sh. not (humans.sh. and animals.sh.)	#9	#1 or #2 or #3 or #4 or #5 or #6 or #7 or #8
#10	8 not 9	#10	animals [mh] not (humans [mh] and animals [mh])
		#11	#9 not #10
Ovid search syntax		**PubMed search syntax**	
.pt. denotes a Publication Type term;		[pt] denotes a Publication Type term;	
.ab. denotes a word in the abstract;		[tiab] denotes a word in the title or abstract;	
.fs. denotes a 'floating' subheading;		[sh] denotes a subheading;	
.sh. denotes a Medical Subject Heading (MeSH) term;		[mh] denotes a Medical Subject Heading (MeSH) term ('exploded');	
.ti. denotes a word in the title.		[mesh: noexp] denotes a Medical Subject Heading (MeSH) term (not 'exploded');	
		[ti] denotes a word in the title.	

(1) Publication Bias

Publication bias (출판 편향)란 저널 편집자들이 negative 연구보다는 positive 연구(통계적으로 유의한 차이를 보인 연구)를 더 선호하기 때문에 positive 연구들은 출판될 가능성이 높게 되고, 따라서 메타분석을 위한 문헌탐색 시 positive 결과를 보인 연구들이 더 많이 파악되어 결과적으로 메타분석결과가 왜곡되게 되는 bias를 의미한다[31]. Publication bias가 존재하는지의 여부를 확인할 수 있는 방법은 여러 가지가 있으나 그 중 가장 자주 사용되는 방법은 funnel plot (깔때기 그림)을 그려보는 방법이다[32]. Funnel plot이란 x-축에는 각 연구에서 추정된 처리효과(예를 들면 OR)를, y-축에는 해당 연구의 정밀성을 나타내는 척도(예를 들면 표본수나 OR의 표준오차 등)를 사용해 산점도(scatter plot)를 그린 것이다[보통의 경우는 x-축 척도의 대칭성을 위해 로그변환한 값들을 사용한다]. '깔때기 그림'이라는 명칭은 연구의 표본수가 커지면 커질수록 실제 효과에 대한 추정이 더 정밀하게 되는 점을 반영해 붙인 이름이다. 즉, 만일 메타분석 결과에 publication bias가 개입되지 않았다면, 소규모 연구들은 연구의 정밀성이 낮기 때문에 이 연구들의 처리효과 추정치들 간에는 변동폭이 심할 것이고 따라서 해당점들은 산점도의 아래 부분에 좌우로 넓게 퍼지게 되는 반면, 대규모 연구들은 정밀성이 높은 연구들이라서 서로 비슷한 크기의 처리효과 추정치들을 제공할 것이고 해당점들 또한 산점도의 윗부분에 서로 모여 있게 되어, 결과적으로 점들이 대칭적이면서 뒤집어진 형태의 깔때기 모양으로 표현될 것이다. 만일 publication bias가 존재하는 경우라면 funnel plot의 점들은 어느 한 쪽으로 치우치게 되어 비대칭적인 형태의 산점도가 그려지게 되는데, 주로 그림의 왼쪽 아랫부분에 빈 공간이 생기게 된다. 왜냐하면 이 부분은 바로 정밀성이 낮으면서 처리효과의 크기도 작은 연구들에 해당하는 점들이 위치하는 부분인데, 처리효과의 크기가 작은 경우 이를 발견해 낼 수 있기에는 너무 적은 표본수를 사용한 연구들은 주로 negative 결과들을 보였을 것이고, 따라서 해당 연구결과들은 발표되지 않았을 가능성이 높아 메타분석을 위한 문헌탐색에서도 발견되지 않았을 것이기 때문이다.

Rosiglitazone 안전성 메타분석에 사용된 42개 논문들 중 rosilitazone군과 대조군 모두에서 MI 발생건수가 보고되지 않았던 4개 논문을 제외한 38개 논문들을 대상으로 MI 발생률에 대한 funnel plot을 그려보면 Fig. 1과 같다. 가로축(x-축)의 0은 rosiglitazone군과 대조군 간 MI 발생 odds에 차이가 없음을 의미하는 값이고, 양수는 rosiglitazone군의 MI 발생 odds가 대조군 에 비해 높음을, 그리고 음수는 그 반대를 의미한다. 가운데 수직 실선은 38개 논문들에 대한 메타분석 결과인 MI 발생 OR에 대한 결합추정치(pooled estimate)를 나타낸다 [Nissen and Wolski (2007)에서 사용되었던 Peto 방법에 의한 결과임]. Funnel plot은 대략적으로 대칭적인 형태를 보여주고 있기 때문에 해당 분석결과에 publication bias가 개입되었을 가능성은 그리 높지 않은 것으로 보인다. 하지만 오른쪽 하단부가 약간 비어 있는 경향

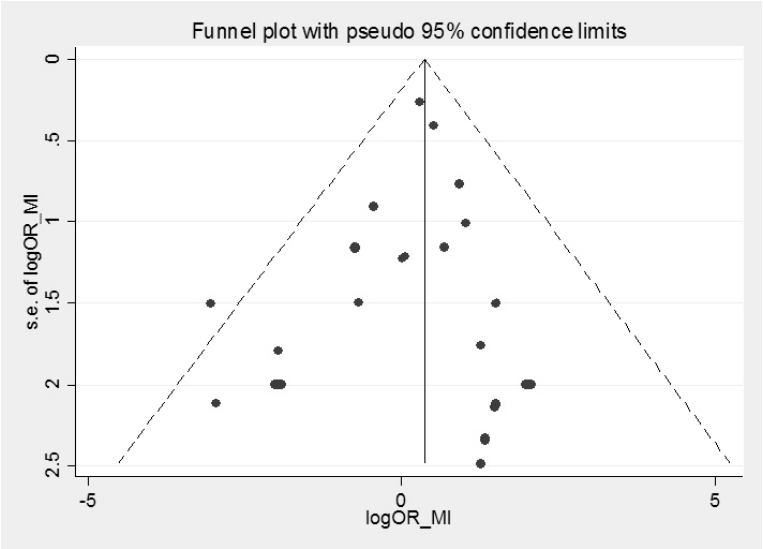

Fig. 1. A funnel plot of log odds ratios (ORs) of the incidence of myocardial infarction between rosiglitazone group and control group versus its standard error (SE). Data used was from the 42 trials displayed in Table 3 of Nissen and Wolski (2007).

을 보이는 것으로 볼 때, 표본수가 적으면서 rosiglitazone군이 대조군에 비해 상대적으로 높은 MI 발생 odds를 보인 연구들이 누락되었을 가능성도 무시할 수는 없을 것 같다(이 메타분석 연구는 주 결과변수로 효과(efficacy)가 아니라 안전정(safety outcome)을 사용한 연구이기 때문에 표본수가 적으면서 rosiglitazone의 risk가 높게 나타나는 부분인 오른쪽 하단부가 비어 있는 경향을 보이게 된 것이다). 하지만 만일 메타분석에 포함되지 않은 해당 연구들이 추가되었다면 결합추정치는 현재보다 더 큰 값으로(즉, rosiglitazone의 위험도가 현재보다는 더 높은 쪽으로) 나타나게 되었을 것이기 때문에, 이 경우의 publication bias는 연구결과를 보다 보수적인 방향으로 유도하는(toward to the null) bias라고 할 수 있을 것이다.

4) 연구의 질 평가 및 자료추출

최종 분석대상 논문들이 선택되면 이들 각 논문으로부터 적절한 정보들을 추출해야 한다. 자료추출을 위해서는 먼저 표준화된 형태로 구성된 자료추출 양식지를 준비해야 하며, 주요 추출 정보들로는 각 연구의 특징에 대한 요약, 환자군의 특성, 처리의 종류, 연구수행 기간, 일차 평가변수 및 기타 평가변수들에 관한 정보 등이다[30]. 이때 추출되는 자료들은 동일한 척도(예를 들면 OR)로 표현될 수 있는 자료들이어야 한다. 따라서 일부 연구들은 이 과정 중에 자료의 부족으로 인해 제외될 수도 있다.

연구의 질에 대한 평가는 매우 중요하다. 연구의 질이란 우리가 보기에 해당 연구가 "실제" 효과를 얼마나 잘 측정했는지에 관한 정도이며, 이러한 측면의 질을 타당도(validity)라 한다. 연구의 타당도는 내적 타당도(internal validity)-해당 연구결과가 원래 연구하고자 했던 상황을 얼마나 정확히 반영하는지에 관한 정도. 즉, bias가 최소화되는 것-와 외적 타당도(external validity)-해당 연구결과가 다른 상황에까지 얼마나 정확히 일반화될 수 있는지에 관한 정도. 즉, 결과의 일반화 가능성-로 구분하며, 메타분석에서 연구의 질 평가란 선택한 논문들에 대해 이러한 타당도를 평가하는 것을 의미한다. 연구의 질 평가에 대한 언급은 Glass (1976)까지 거슬러 올라간다[16]. 메타분석 시 연구의 질에 대한 평가가 반드시 이루어져야 하는 이유는 분석에 사용된 연구들이 실제로는 질이 낮은 연구들임에도 불구하고 분석결과는 마치 높은 수준의 연구를 통해 얻어진 결과인 것처럼 제시되거나 보일 수 있기 때문이며, 이보다 더 중요한 문제는 질이 낮은 연구결과들을 결합하게 되면 bias된, 따라서 결과가 왜곡(misleading)되는, 결합추정치가 얻어지게 되기 때문이다. 'Garbage in, garbage out', 즉, 아무리 정교한 분석방법을 사용해도 질이 낮은 자료의 한계점은 극복할 수 없다. 메타분석을 수행할 때 가장 간과할 수 없는 형태의 bias가 바로 이 연구의 질 평가와 관련된 bias이며[33], 모의실험 결과, 연구의 질에 주로 영향을 미치는 것은 부적절한 방법론의 사용, 그 중에서도 특히 무작위 배정에 대한 은폐(random allocation concealment)가 제대로 이루어지지 않았을 때 가장 bias(주로 처리효과가 과장되는 방향으로 bias)가 큰 결과가 초래되는 것으로 나타났다[34].

연구의 질 평가에서 고려해야 될 요인들로는 무작위화 실시여부, 배정 은폐에 관한 언급여부, 이중눈가림 실시여부, 연구철회(withdrawal) 및 탈락(drop-out)에 대한 언급여부, 그리고 적절한 통계분석 사용여부 등이다[35]. 연구의 질 평가는 'guilty unless proven innocent', 즉, 실시했다고 언급되어있지 않는 이상 실시하지 않은 것으로 간주하는 유죄추정의 원칙을 사용하는 것이 일반적이다. 연구의 질을 평가하

는 방법으로는 크게 두 종류의 접근방법이 사용된다. 하나는 복합적인 형태의 점수화된 평가도구를 사용하는 척도적 접근방법(scale approach)이고 다른 하나는 연구특성을 개별적으로 평가하는 항목별 접근방법(component approach)이다. 질 평가를 위해 개발된 척도들은 많지만, 이들 중에는 실제로 연구의 질을 측정하는 것이 아니라 논문내용의 적절성이나 연구결과의 일반화와 관련된 외생요인들을 평가하는 데 초점이 맞추어져 있어 타당성에 문제가 있는 것들도 있다[36]. 또한 이러한 척도를 사용하는 방법은 각 연구의 전반적인 질에 관한 정도를 수치화 할 수 있다는 장점은 있으나 어떤 척도를 사용하느냐에 따라 결과가 많이 달라진다는 단점도 있다[37,38]. 그러나 그 간편성 때문에 요소적 접근방법보다 더 많이 사용되는 것이 사실이다. 현재 주로 사용되고 있는 질 평가 척도로는 Chalmers scale[39] 및 Jadad scale[40] 등을 들 수 있는데, Chalmers scale은 연구의 내적, 외적 타당도 및 자료분석 측면 등을 평가하는 전체 30개 문항(100점 만점)으로 이루어져 있으며, Jadad scale은 연구의 내적 타당도를 평가하는 5개 문항, 즉, 무작위화가 실시되었다고 기술되었는가? 연구대상자들의 처리군 할당에 사용된 일련번호들은 적절하게 생성되었다고 기술되었는가? 이중눈가림이 실시되었다고 기술되었는가? 대조군에 사용된 처치(예를 들면 위약)는 처리군의 처치와 구별될 수 없었다고 묘사되었는가? 각 집단 별로 추적실패 또는 분석에서 제외된 환자의 수 등 손실내용(attrition) 및 그 이유가 기술되어 있는가? 등으로 구성된 5점 만점의 척도이다. Jadad scale이 가장 많이 사용되는 척도이긴 하지만, 연구의 실제 방법론적인 측면보다는 작성된 논문의 내용에 관한 질을 평가하는 척도라는 점, 그리고 무작위 배정과 관련된 은폐 여부보다는 무작위 배정 일련번호의 생성 여부에 초점이 맞추어져 있다는 것에 대한 비판도 있다[25,37]. 이에 반해 항목별 접근방법은 처리효과 추정치와 관련된 연구 방법론적 질에 대한 평가로 처리의 배정 은폐에 관한 측면, 무작위 배정 일련번호의 생성에 관한 측면, 모든 무작위 배정된 환자들의 분석 포함여부에 관한 측면, 그리고 이중눈가림에 관한 측면 등 4개 측면에 대한 평가를 각 메타분석의 연구주제에 알맞게 개별적으로 평가하는 방법이다[34]. 최근에는 되도록이면 이 방법을 사용할 것을 추천하고 있지만 연구자가 각자의 메타분석 연구내용에 알맞게 평가측면들을 정하는 게 좋다[28,38]. 결론적으로 연구의 질 평가 시 어떤 접근방법을 사용할 것인가, 그리고 질 평가 내용에 대해서는 단순제시만 할 것인가, 아니면 그 결과를 자료분석 단계에서도 이용할 것인가 등에 대해서는 아직도 논란이 많긴 하지만, 선택된 연구들에 대한 질 평가는 반드시 실시되어야 한다는 점에는 이론의 여지가 없다고 할 수 있다.

자료를 추출할 때나 연구의 질을 평가할 때에는 정확성을 유지하기 위해 두 명의 연구자가 각각 독립적으로 실시해야 한다. 또한 평가결과에 대한 평가자들 간의 일치도를 계산해야 하고(예: 카파 통계량), 불일치 결과가 발생했을 때는 어떻게 해결했는지에 대해서도 기록해야 한다[30]. 논문의 저자명 및 이를 확인할 수 있는 다른 정보들에 대해 평가자들을 눈가림하는 것 역시 bias를 줄일 수 있는 방법 중 하나이다[28].

5) 통계분석 및 결과 제시

메타분석의 마지막 단계는 적절한 통계적 방법을 사용해 연구결과들을 결합하는 작업이다. 자료의 결합은 관심 변수에 관한 각 연구의 개별 환자 자료(individual patient data, IPD)를 사용해 결합할 수도 있고(물론 이를 위해서는 각 연구의 환자 자료를 얻어야 함), 각 연구의 해당 변수에 대한 요약자료를 사용할 수도 있다(IPD 메타분석에 대해서는 참고문헌 41 및 42 등을 참조). 요약자료를 사용해 처리 간 효과 차이에 대한 메타분석을 하기 위해서는 각 연구에 대해서 기본적으로 다음과 같은 두 가지 정보가 필요하다.

① 처리효과 차이에 대한 동일한 형태의 측정값: 예를 들면, 교차비(OR), 위험비(RR) 또는 평균차이(MD) 등

② 해당 처리효과에 대한 표준오차(또는 분산)

메타분석에서 각 연구의 효과 차이들을 결합할 때 사용하는 주 개념은 각 연구의 표본수 크기에 근거한 가중평균을 사용하는 것이다. 즉, 소규모 연구는 대규모 연구에 비해 우연한 차이로 인한 영향을 더 많이 받기 때문에 결합 시 대규모 연구에 상대적으로 더 많은 가중치를 부여하고자 하는 것이다[30]. 결합효과 및 해당 신뢰구간을 추정하고 결합효과에 대한 가설검정을 실시하기 위해 메타분석에서 주로 사용되는 통계적 모형으로는 고정효과모형(fixed-effects model)과 변량효과모형(random-effects model)이 있다. 이들 두 모형의 차이는 각 논문에서 얻어진 요약자료가 어떤 분포에서 얻어진 자료인지에 대한 통계적인 관점의 차이라 할 수 있다[43].

(1) 고정효과모형(Fixed-effects Model)

이 통계모형은 각 연구들이 보고자 했던 처리효과는 사실 같은 것인데 연구결과가 서로 다르게 나타난 이유는 연구들의 표본추출오차 때문이라고 가정하는 모형이다. Fig. 2(a)는 이를 좀 더 알기 쉽게 그림으로 표현한 것이다. 즉, 각 연구에서 관찰된 처리효과들은 하나의 공통된 실제 처리효과(common true treatment effect)를 중심으로 무작위로 나타난 결과들이라는 가정이다. 따라서 이 모형은 각 연구들의 처리효과들 간에 보이는 변동의 크기가 단순한 표본추출오차에 기인한 정도라고 믿어지는 경우, 즉, 각 연구들의 연구설계나 방법론 등이 서로 유사한 경우에 사용되는 모형이다. 이 모형을 사용해 결합효과(결합추정치, pooled estimate)를 추정하는 통계적 방법들로는 역분산가중 추정법(inverse variance

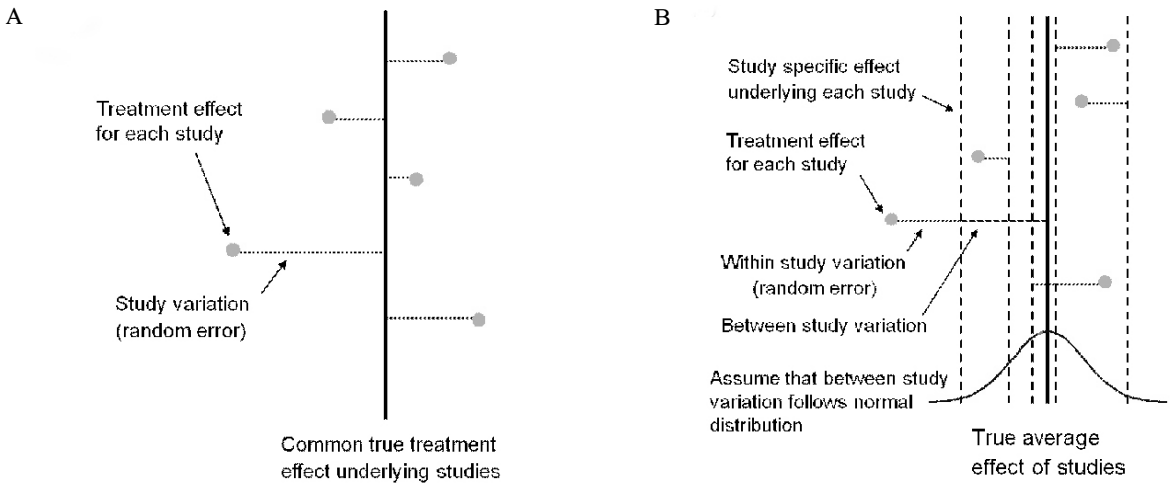

Fig. 2. A graphical representation of the concepts of (a) a fixed-effects model and (b) a random-effects model. A fixed-effects model assumes that there is a common 'true' treatment effect underlying studies and that each study result will vary randomly around this true effect (i.e., treatment effect for each study = common true treatment effect + study variation that called a random error). A random-effects model assumes that there is different underlying effect for each study and that there is also an additional source of variation between studies which reflects the amount of heterogeneity between them (i.e., treatment effect for each study = true average effect which follows a normal distribution + within study variation that called a random error).

weighted estimation method, IVW), 멘틀-헨젤 추정법 (Mantel-Haenszel estimation method, MH), 피토 추정법 (Peto estimation method), 그리고 최대 가능도 추정법 (maximum likelihood estimation method, ML) 등이 있다. 이들 중 가장 많이 사용되는 방법은 IVW 추정방법이다. 이 방법은 가중치로 각 연구의 정밀성을 나타내는 효과차이에 대한 분산의 역수를 사용해 연구결과들을 결합하는 방법이다. IVW 추정방법을 OR들을 결합하는 경우에 대해 간단히 소개하면 Table 2와 같다. 각 추정방법들은 나름대로의 장단점을 가지고 있다[25,44]. IVW 추정법은 결합하고자 하는 연구들의 수는 적지만 각 연구들이 표본수가 큰 연구들인 경우에 효과적인 방법이다. MH 추정법은 표본수가 작은 연구들이라 할지라도 그러한 연구들의 수가 많을 때 효과적인 방법으로, 처리효과들을 결합할 때 IVW 방법처럼 각 연구의 OR값에 대해 로그변환을 하지 않고 OR값 자체를 사용한다. Peto 추정법은 계산하기가 쉽고, 결과(event) 수가 적거나 발생률(event rate)이 낮은 경우, 그리고 처리군이나 대조군 중 하나에서 결과가 발생하지 않은 연구들이 있는 경우(이 경우 IVW 방법이나 MH 방법은 해당 연구를 제외하고 결합하게 된다), 즉, 빈 칸(zero cell)이 있는 경우에 효과적인 방법이다. 그러나 Peto 방법은 두 처리집단 간 표본수에 불균형이 심한 연구들이 많으면 bias된 OR값(주로 1에서 멀어지는 쪽으로) 및 표준오차를 제공하게 되고, 반면에 true OR가 큰 경우에는 이를 과소추정(under-estimate)해 주는 경향이 있기 때문에 비판을 많이 받는 방법이다. ML 방법은 연구들의 표본수가 작을 때 매우 효율적인 방법이고, 여러 통계적인 장점을 많이 가지고 있는 방법이긴 하나, 계산하기가 어렵고, 표본수가 클 때에는(각 칸의 값들이 5 이상인 경우) MH 방법과 거의 동일한 결과를 제공해준다. 범주형 자료의 경우, MH 방법이나 Peto 방법은 2 × 2 분할표의 경우에만 사용할 수 있지만, IVW 방법이나 ML 방법은 일반적인 경우에도 사용할 수 있다. 메타분석 시 어떤 방법을 선택할 것인가 하는 것도 중요하지만 그에 못지않게 사용한 방법을 명확히 기록해 주는 것도 중요하다.

Rosiglitazone safety 연구에서는 각 연구들의 event 수가 낮고, 빈 칸들도 있기 때문에 Peto 방법을 사용해 메타분석을 실시하였다. Nissen과 Wolski (2007)의 Table 3 (pp. 2466~2467) 자료를 사용해 Peto 방법에 기초한 고정효과모형 메타분석결과를 STATA 프로그램으로 재현해보면 Fig. 3과 같다. 분석결과, control 군에 대한 rosiglitazone 군의 MI 발생 OR 결합추정치 및 해당 95% 신뢰구간은 1.43 (1.03~1.98)이었고, CVD 사망률의 경우에는 1.64 (0.98~2.74)이었다. 즉, rosiglitazone 치료는 대조군(control treatment)에 비해 MI 발생 odds를 43% 정도(신뢰범위 3%~98%) 증가시키는 것으로 나타났다. 참고로 이 자료에 대해 MH 방법을 사용하면 MI 발생률은 1.28 (0.95~1.72), CVD 사망률은 1.33 (0.83~2.13)인 반면, IVW 방법을 적용하면 MI 발생률은 1.29 (0.94~1.76), CVD 사망률은 1.31 (0.80~2.13)으로 계산된다.

메타분석 결과를 요약해서 제시하는 방법 중 하나로 숲그림(forest plot)이 많이 사용된다. 이 그림은 각 연구들의 처리효과에 대한 추정치들 및 해당 신뢰구간, 그리고 메타분석 결과로 얻어진 결합추정치와 그 신뢰구간을 하나의 그림으로 표현한 것으로, 마치 나무들이 모여 있는 숲처럼 보인

Table 2. A fixed-effects model meta-analysis using inverse-variance weighted (IVW) estimation method to obtain a pooled odds ratio (OR) estimate and its 95% confidence interval (CI) based on 2 × 2 tables comparing binary outcome between treatment and control groups

자료의 형태 - k개의 연구가 있으며, 각 연구들은 다음과 같은 2 × 2표 형태의 자료가 있다고 가정하자.

Treatment	Event		
Yes	No	Total	
A			
B	a		
c	b		
d	a + b		
c + d			
Total	a + c	b + d	n

필요 가정들 - 모든 연구들은 동일한 크기의 실제 OR를 가지고 있다.
- 각 연구에서 얻어진 OR들이 서로 다르게 관찰된 이유는 각 연구 내에서 표본추출 변동(within-study variation, 즉, random variation)이 발생했기 때문이다.

계산 방법 [1] 각 연구의 OR값에 대해 자연로그를 취함:

예를 들어 i번째 연구의 경우, $\ln(OR_i) = \ln\left(\frac{a \times d}{b \times c}\right)$

[2] 각 연구의 ln(OR)값에 대한 표준오차(SE)를 구함:

예를 들어 i 번째 연구의 ln(OR)값에 대한 표준오차는 $SE_i = \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$

[3] 각 연구에 대한고정효과모형의 가중치를 계산함:

$$w_i = \frac{1}{SE_i^2}$$

[4] ln(OR)에 관한 결합추정치(pooled estimate)를 구함:

$$\ln\left(OR_{pooled}\right) = \frac{\sum_{i=1}^{k} w_i \ln(OR_i)}{\sum_{i=1}^{k} w_i}$$

[5] ln(OR) 결합추정치에 대한 표준오차(SE)를 구함:

$$SE_{pooled} = \frac{1}{\sqrt{\sum_{i=1}^{k} w_i}}$$

[6] OR에 관한 결합추정치인 OR_{pooled}를 구하기 위해 위 [4]번의 ln(OR) 결합추정치를 역-로그 변환함:

$$OR_{pooled} = e^{\ln\left(OR_{pooled}\right)}$$

[7] 결합추정치 OR_{pooled}에 대한 95% 신뢰구간을 구하기 위해 위 [4], [5]번 결과를 이용함:

$$95\%\ CI\ (OR_{pooled}) = e^{\ln\left(OR_{pooled}\right) \pm 1.96 \times SE_{pooled}}$$

결과의 해석 - OR_{pooled}는 고정효과모형을사용해서 얻어진 결합처리효과를 나타낸다.
- 이 OR_{pooled}에 대한 95% 신뢰구간은 실제 처리효과가 존재할 것으로 믿어지는 범위를 나타낸다.
- 만일 이 신뢰구간이 1을 포함하고 있지 않다면 처리 A의 효과와 처리 B의 효과는 서로 다르다고 본다.
- 만일 이 신뢰구간이 1을 포함하고 있으면 두 처리의 효과가 서로 다르다는 통계적인 증거는 없다고 본다.

다고 해서 forest plot이라는 명칭이 붙었다.

Rosiglitazone safety 메타분석의 MI 발생률에 대한 forest plot은 Fig. 4와 같다. 이 그림에 사용한 기호들은 Fig. 4에 자세히 설명해 두었다. 이 forest plot을 보면 각 연구결과들은 두 군 간에 통계적인 차이가 없음에도 불구하고 메타분석 결과는 통계적으로 차이가 있는 것으로 나타났음을 볼 수 있다. 그 이유는 메타분석을 통해 전체 표본수가 커졌기 때문에 두 군 간 차이를 탐지해 낼 수 있는 통계적 검정력이 증가되었기 때문이다. 한편 가중치의 크기를 볼 때 42개 논문들 중 두 개의 대규모 RCT 연구(DREAM trial 및 ADOPT trial)가 전체 가중치의 약 58%를 차지하고 있음을 볼 수 있다.

(2) 변량효과모형(Random-effects Model)

변량효과모형은 각 연구에서 보고자 했던 실제 처리효과가 각각 다르다고 가정하는 모형이다. 이는 곧 각 연구설계

Table 3. A random-effects model meta-analysis using DerSimonian-Laired (DL) estimation method to obtain a pooled odds ratio (OR) estimate and its 95% confidence interval (CI) based on 2 × 2 tables comparing binary outcome between treatment and control groups

자료의 형태	- Table 1과 동일한 상황을 가정하자.
필요 가정들	- 각 연구의 실제 OR값들은 서로 다르다. - 관찰된 OR값들이 서로 다른 이유는 각 연구 내 표본추출 변동(within-study variation, 즉, random variation)뿐만 아니라 연구 간 변동(between-study variation), 즉, 연구 간 이질성(heterogeneity between studies) 때문이기도 하다.
계산 방법	[1] 연구들 간의 이질성 존재 여부에 관한 검정을 실시함: - 이질성 여부를 파악하기 위해 자주 사용되는 방법으로는 Cochran의 Q-검정을 실시하는 방법과 Higgins의 I^2-값을 이용하는 방법이 있음(자세한 내용은 본문 내 설명 참조). 여기서는 Cochran의 Q-검정에 대해 설명함. - 귀무가설: k개의 실제 OR들은 모두 동일한 값들이다. - 검정통계량: $Q = \sum_{i=1}^{k} w_i \left(OR_i - OR_F\right)$ - 여기서 wi는 고정효과모형에서 계산된 i번째 연구의 가중치를, OR_i는 i번째 연구에서 관찰된 OR값을, 그리고 OR_F는 고정효과모형을 사용해 얻어진 OR에 대한 결합추정치를 나타내며, 해당 값들에 대한공식은 Table 1에 주어져 있다. [2] 연구 간 변동에 관한추정치를 계산함: $\tau = \max\left(0,\ \dfrac{Q-(k-1)}{\sum_{i=1}^{k} w_i - \left(\sum_{i=1}^{k} w_i^2 \Big/ \sum_{i=1}^{k} w_i\right)}\right)$ [3] 각 연구에 대한 변량효과모형의 가중치를 계산함: $w_i^* = \dfrac{1}{\dfrac{1}{w_i} + \tau}$ [4] ln(OR)에 관한 결합추정치(pooled estimate)를 구함: $\ln\left(OR_{pooled}\right) = \dfrac{\sum_{i=1}^{k} w_i^* \ln(OR_i)}{\sum_{i=1}^{k} w_i^*}$ [5] ln(OR) 결합추정치에 대한 표준오차(SE)를 구함: $SE_{pooled} = \dfrac{1}{\sqrt{\sum_{i=1}^{k} w_i^*}}$ [6] OR에 관한 결합추정치를 구하기 위해 위 [4]번의 ln(OR) 결합추정치를 역-로그 변환함: $OR_{pooled} = e^{\ln\left(OR_{pooled}\right)}$ [7] 결합추정치 OR_{pooled}에 대한 95% 신뢰구간을 구하기 위해 위 [4], [5]번 결과를 이용함: $95\%\ CI\ (OR_{pooled}) = e^{\ln\left(OR_{pooled}\right) \pm 1.96 \times SE_{pooled}}$
결과의 해석	- 만일 Q = k - 1이면(또는 이에 해당하는 *P값* < 0.1이면) 연구들 간에 통계적 이질성(statistical heterogeneity)에 관한 증거가 있다고 본다. 만일 Q < k - 1이면(또는 *P값* > 0.1이면) OR값들이 서로 다르다는 증거는 없다고 본다. - OR_{pooled}는 변량효과모형을 사용해서 얻어진 결합처리효과를 나타낸다. - 이 OR_{pooled}에 대한 95% 신뢰구간은 실제 처리효과가 존재할 것으로 믿어지는 범위를 나타낸다. - 만일 이 신뢰구간이 1을 포함하고 있지 않다면 처리 A의 효과는 처리B의 효과와 다르다고 본다.

들이 서로 다른 특성을 가지고 있기 때문에(예를 들면 각 연구에서 사용된 연구모집단, 지역, 또는 dose 등이 다르기 때문에) 이들 메타분석에 포함되는 연구들이 모두 동일한 처리효과를 추구하고 있다고 볼 필요는 없다는 관점이다. 이 모형에 대한 개념을 그림으로 표현하면 Fig. 2(b)와 같다. 즉, 각 연구들은 어떤 평균적인 처리효과(true average treatment effect)를 중심으로 퍼져있는(주로 정규분포를 가정한다) 모집단 내 연구들로부터 무작위로 추출된 연구들이라고 간주하고, 따라서 각 연구의 처리효과들 간에 관찰되는 변동은 표본추출변동(sampling variation, within-study variation, random variation)과 각 연구들 간의 변동(between-study variation)이 함께 표현된 것이라고 가정한다. 그러므로 변량효과모형 메타분석에서는 자료결합 시 이 연구들 간의 변동과 각 연구 내의 무작위 변동을 동시에 고려한 가중치를

```
. metan r_mi r_mi_no c_mi c_mi_no, or peto chi2 label(namevar=study_id)

          Study        |     OR    [95% Conf. Interval]     % Weight
-----------------------+-------------------------------------------------
01:49653/011           |   4.463       0.234    85.244          1.22
02:49653/020           |   1.058       0.098    11.466          1.87
03:49653/024           |   0.139       0.004     4.663          0.86
04:49653/093           |   0.052       0.001     3.280          0.62
05:49653/094           |   4.482       0.070   286.492          0.61
06:100684              |   0.147       0.003     7.455          0.69
07:49653/143           |   7.575       0.150   381.845          0.69
08:49653/211           |   2.504       0.557    11.246          4.71
09:49653/284           |   7.428       0.147   374.345          0.69
10:712753/008          |   4.373       0.066   289.868          0.60
11:AVM100264           |   0.139       0.003     7.006          0.69
12:BRL49653C/185       |   3.504       0.110   111.281          0.89
13:BRL49653/334        |   1.959       0.203    18.915          2.07
14:BRL49653/347        |   4.525       0.240    85.195          1.23
15:49653/015           |   1.003       0.091    11.090          1.84
16:49653/079           |   0.497       0.027     9.257          1.24
17:49653/080           |   0.485       0.050     4.719          2.05
18:49653/082           |   4.524       0.239    85.581          1.23
19:49653/085           |   2.768       0.386    19.866          2.74
21:49653/097           |   0.133       0.003     6.709          0.69
22:49653/125           |   0.134       0.003     6.743          0.69
23:49653/127           |   7.658       0.152   386.158          0.69
24:49653/128           |   7.202       0.143   363.078          0.69
25:49653/134           |   0.048       0.003     0.917          1.22
26:49653/135           |   0.636       0.108     3.732          3.40
27:49653/136           |   7.144       0.142   360.223          0.69
28:49653/145           |   7.749       0.154   390.961          0.69
29:49653/147           |   7.306       0.145   368.247          0.69
30:49653/162           |   7.567       0.150   381.461          0.69
32:49653/330           |   3.750       0.039   361.055          0.51
34:49653/137           |   0.463       0.048     4.489          2.06
35:SB71253/002         |   7.187       0.143   362.324          0.69
36:SB71253/003         |   7.932       0.157   400.651          0.69
37:SB71253/007         |   4.439       0.069   287.643          0.61
39:49653/132           |   3.502       0.027   461.091          0.45
40:AVA100193           |   3.724       0.038   367.954          0.50
41:DREAM               |   1.652       0.741     3.683         16.53
42:ADOPT               |   1.329       0.800     2.209         41.25
20:49653/095           |  (Excluded)
31:49653/234           |  (Excluded)
33:49653/331           |  (Excluded)
38:SB71253/009         |  (Excluded)
-----------------------+-------------------------------------------------
Peto pooled OR         |   1.428       1.031     1.979        100.00
-----------------------+-------------------------------------------------

  Heterogeneity chi-squared =    29.36 (d.f. = 37) p = 0.810
  I-squared (variation in OR attributable to heterogeneity) =       0.0%

  Test of OR=1: chi-squared =   4.59 (d.f. = 1) p =  0.032
```

Fig. 3. A sample result of meta-analysis for the odds ratio (OR) of myocardial infarction (MI) between rosiglitazone group and control group using STATA software based on the data from the Table 3 of Nissen and Wolski (2007). Peto estimation method is used to combine the results of 38 trials. Four trials are excluded in the meta-analysis due to zero events observed in both groups.

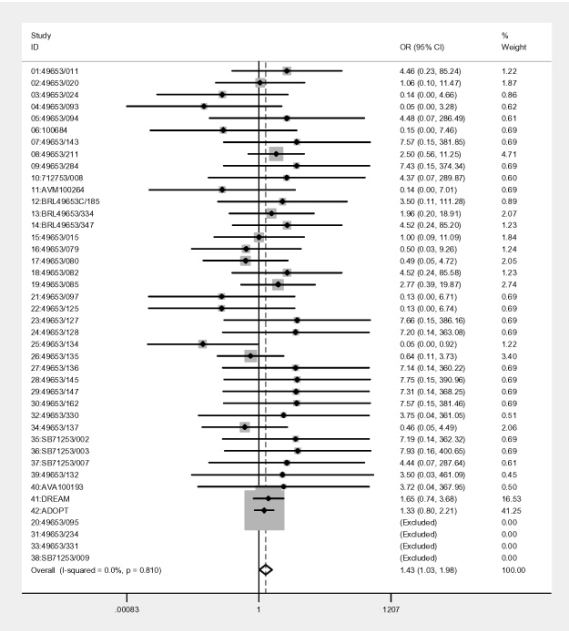

Fig. 4. A forest plot and the combined result of meta-analysis for the odds ratio (OR) of myocardial infarction (MI) between rosiglitazone group and control group. A fixed-effects model with Peto estimation method is used. Each of blocks represents the OR for each trial and the horizontal line indicates its 95% confidence interval (CI). The size of each block is approximately proportional to the statistical weight of the trial used in the meta-analysis. The diamond represents the pooled estimate and its 95% CI. The solid vertical line indicates no difference in the odds of MI between two groups, while the dashed vertical line represents the pooled effect. The left side of the solid vertical line represents lower odds of MI for rosiglitazone than for control, while the right side represents higher odds. Four trials are excluded in the meta-analysis due to zero events observed in both groups.

사용하게 된다. 만일 연구들 간에 이질성이 전혀 없다면 변량효과모형은 고정효과모형과 동일한 결과를 제공한다. 하지만 연구들 간에 이질성이 존재한다면 변량효과모형은 고정효과모형에 비해 연구들 간에 더 많은 변동을 인정하기 때문에, 결합추정치에 대해서 더 넓은 신뢰구간을 제공하게 된다. 또한 변량효과모형은 자료결합 시, 고정효과모형에 비해, 소규모 연구에 상대적으로 더 많은 가중치를 부여한다. 따라서 만일 publication bias가 존재하는 경우라면 두 모형의 결과는 서로 달라진다[24,45]. 사실 연구들 간의 변동을 인정하는 변량효과모형은 메타분석에 포함되는 연구들이 어떤 가상적인 연구들로 구성된 모집단으로부터 무작위로 추출된 연구들이라고 가정하는 것이기 때문에, 자료를 이런 방식으로 접근하는 관점은 개념적으로 문제가 있다고 비판하는 사람도 있으나[24], 이제는 널리 받아들여지는 통계적인 개념이 되었다. 따라서 연구결과들 간에 나타난 이질성 정도가 단순히 해당 연구들의 특성(즉, 무작위 변동) 때문인 것으로만 간주할 수 없을 정도로 너무 크다면 변량효과모형을 사용할 필요가 있다. 변량효과모형을 사용해 결합추정치를 계산하는 통계적 방법으로는 가중최소제곱 추정법(weighted least squares estimation method, WLS), 비가중최소제곱 추정법(unweighted least squared estimation method, UWLS), 최대가능도 추정법(maximum likelihood estimation method, ML), 제한최대가능도 추정법(restricted maximum likelihood estimation method, REML) 등이 있다. 이중 WLS 추정법이 결과의 안정성이나 계산상의 간편성 때문에 가장 널리 사용되며, 이 방법을 개발한 연구자들의 이름을 따서 보통 DerSimonian-Laired (D-L) 추정법이라 부른다[24]. Table 3은 OR들을 결합하는 경우를 예로 들어 D-L 추정방법을 간단히 소개한 표이다.

고정효과모형의 분석결과와 비교하기 위해 rosiglitazone 안전성 메타분석 연구자료에 대한 변량효과모형 메타분석을 실시해본 결과, D-L 추정법에 기초한 MI 발생률의 결합추정치는 1.29 (0.94~1.76), 그리고 CVD 사망률의 경우에는 1.31 (0.81~2.13)이다. 이 결과들은 고정효과모형의 MH 추정법을 사용한 결과와 비교해 볼 때 결합추정치의 크기는 비슷하지만 신뢰구간은 더 넓어진 것을 볼 수 있다.

(3) 통계적 이질성의 존재여부에 대한 평가

연구들 간의 통계적 이질성(statistical heterogeneity)이란 각 연구결과에 관한 요약자료(처리효과에 대한 측정값 및 해당 신뢰구간)의 크기들이 통계적으로 서로 다른 것을 의미한다. 결합처리효과에 대한 고정효과모형의 결과와 변량효과모형의 결과가 서로 같다면 연구들 간에는 통계적으로 유의한 이질성이 존재하지는 않는다는 의미이지만, 만일 연구들 간에 통계적인 이질성이 많이 존재한다면 두 모형의 메타분석 결과는 상당한 차이를 보이게 된다. 이런 경우에는 더 보수적인 결과, 즉, 일반적으로 변량효과모형에 의한 분석결과를 제시해야 된다.

연구결과들 간의 통계적 이질성 존재여부에 관한 평가는 Cochran의 Q-검정[44,46]과 Higgins의 I^2-통계량[47,48]을 이용하는 방법이 주로 사용된다. Cochran의 Q-검정은 통계적 검정을 통한 *P*값을 제시해 주긴 하지만(이때 귀무가설은 '연구결과들이 동질적이다'라는 것이며, *P*값이 낮으면 연구들 간에 통계적인 이질성이 존재한다는 의미이다) 메타분석의 경우 Q-검정에 사용되는 연구의 수가 적기 때문에 실제 차이를 발견해 낼 수 있는 통계적인 검정력이 낮아지게 된다. 따라서 실제로 연구결과들 간에 이질성이 존재하는 경우라 할지라도 일반적인 유의수준 0.05 하에서 Q-통계량은 유의하지 않은 결과를 보일 가능성이 높으며, 이에 따라 메타분석에서는 Fleiss (1986)가 제안한대로 유의수준 0.1을 사용하는 것이 보통이다[49]. Higgins의 I^2-통계량은, Q-통계량과는 달리, 연구의 수나 결과변수의 형태 또는 처리효과의 종류(예를 들면 OR) 등에 영향을 받지 않도록 만든 지표로, 연구들 간의 이질성 정도가 메타분석 결과에 미치는 영향력의 크기를 정량화해 주는 척도이다[21]. I^2-통계량은 연구결과들 간의 전체 변동량 중 이질성으로 인한 변동량이 차지하는 비율로 정의되며, $I^2 = [(Q - df) / Q]\ 100\ (\%)$으로 계산된다(여기서 df = k - 1, k = 사용된 연구의 수). 이 값은 0%에서 100% 사이의 값을 가지게 되며, 만일 연구들 간에 관찰되는 이질성이 없으면 0%가 된다. 계산 상 음수가 나오는 경우 역시 0%로 정의한다. 이 값에 대한 일반적인 사용기준으로는 $I^2 < 25\%$이면 통계적 이질성이 낮은 것으로, $25\% < I^2 < 75\%$이면 중간 정도의 이질성이 있는 것으로, 그리고 $I^2 > 75\%$이면 이질성 정도가 심한 것으로 간주하며[48], 모형 선택 시의 기준 값으로는 $I^2 = 50\%$가 주로 이용된다.

Rosiglitazone 안전성 메타분석의 MI 발생률에 대한 자료의 경우, Q = 29.36, df = 37, P = 0.810이므로 $I^2 = 0\%$로 계산되었고(Fig. 3 하단부 참조), 이 결과들은 고정효과모형을 사용하는 것이 타당함을 지지한다고 볼 수 있다.

(4) 부집단분석 - 이질성의 원인 탐색

만일 연구결과들 간에 통계적인 이질성이 존재한다는 증거가 있으면 메타분석은 조심스럽게 진행되어야 하며, 분석방법 또한 그에 맞게 조정되어야 하고, 이질성이 나타나게 된 원인도 조사되어야 한다. 통계적 이질성이 발견되지 않았다면 고정효과모형을 사용하는 것이 좋다. 만일 이질성의 정도가 약한 경우라면 변량효과모형에 기초한 메타분석이 주로 사용된다. 그러나 통계적 이질성이 존재하는 것으로 판단된다면 과연 이들 연구들을 결합하는 것이 의미가 있는가? 이러한 연구들을 결합하는 것은 사과에 오렌지를, 그리고 여기에 레몬을 섞는 것이나 다를 바 없다는 비판도 있으며[50],

연구들을 결합해도 좋은 이질성의 크기는 얼마까지인가에 대한 명확한 기준도 없다. 그러나 변량효과모형은 연구결과들 간의 이질성을 해결하기 위한 방법이 아니라 단지 각 연구 내의 무작위 변동 외에 연구결과들 간의 변동을 추가적으로 고려한 통계적인 모형일 뿐이다. 따라서 이보다 중요한 것은 이러한 통계적인 이질성이 나타나게 된 원인들을 파악하는 것이라 할 수 있다. 이질성의 원인들을 파악하기 위해 주로 사용되는 방법으로는 부집단 분석(subgroup analysis)을 실시하는 방법[51], 통계적 방법인 메타회귀분석(meta-regression analysis)을 사용하는 방법[52], 그리고 그래프를 이용하는 방법 등이 있다. 부집단 분석은 환자군이나 연구방법의 특성에 따라 대상 연구들을 동질적인 세부집단으로 분류해서 분석하는 것을 말한다. 부집단 분석의 목적은 연구들 간의 통계적인 이질성이 임상적인 이질성(clinical heterogeneity) - 예를 들면, 연구들 간 환자 포함기준, 환자들의 기저질환 정도, 연령대 및 성별, 진단기준, 용량 또는 약제 복용 스케줄, 처리의 종류, 결과변수에 대한 정의, 추적기간, 연구 지역 및 연구 상황 등의 차이 여부-이나 방법론적인 이질성(methodological heterogeneity) - 예를 들면, 연구설계의 차이(병행설계 혹은 교차설계), 무작위 방법의 차이(개인별 혹은 집락 무작위화), 연구 질의 차이(무작위배정 여부나 이중눈가림 실시 여부 등), 분석방법의 차이(ITT 분석결과 혹은 PP 분석결과) 등에 의해 기인한 것인지를 파악하고자 하는 것이다[53]. 그러나 부집단 분석은 사용되는 연구의 수가 더 작아져 검정력이 굉장히 낮아지기 때문에 해석에 더욱 주의를 기울여야 한다. 즉, 부집단 분석결과는, 연구주제에 대한 답을 내리기 위한 것이 아니라, 이질성의 원인에 대한 탐색 및 추가 연구를 위한 새로운 가설을 구축하기 위한 용도 정도로만 사용하는 것이 좋다.

참고로 이질성 탐색을 위해 자주 사용되는 그래프로는 Forest plot이나 L'Abb plot[26,54], 혹은 Galbraith plot[55] 등이 있다[25,44].

(5) 민감도분석

민감도 분석(sensitivity analysis)은 메타분석 결과가 이를 수행하는 과정 중 각 단계에서 행해진 주요 결정사항들이나 가정들에 대해 쉽게 변화되는지를 평가하고자 하는 것으로[28], 주 분석에 이어 부수적으로 이루어지는 추가분석에 해당한다. 왜냐하면 만일 분석결과가 분석방법이나 분석과정 중에 내려진 여러 의사결정들에 대해 민감하지 않다면(robust) 우리는 이 결과에 대해 한층 더 자신감을 가질 수 있기 때문이다. 민감도 분석을 위해 코크란 가이드라인에서는 연구 포함기준(환자군, 중재, 결과변수의 정의 등)을 변화시키면서 재분석 실시, 포함기준 만족여부가 불분명한 연구들을 포함/제외시켜 보면서 그 효과를 검사, 공식적인 상호심사과정을 거치지 않은 논문들이나 미출판된 연구들을 포함/제외시켜 보면서 재분석 실시, 질이 낮은 연구들이 분석결과에 미친 영향은 얼마나 되는지 그 정도를 검토, 추출된 자료들 중 신뢰성이 낮다고 판단되는 자료에 대한 재검토, publication bias에 대한 평가, 결측값이 존재하는 경우에는 이를 적절한 범위 내의 값들로 대체시키면서 재분석 실시, 결과예측이 어느 정도 가능할 만큼 실시된 외부연구가

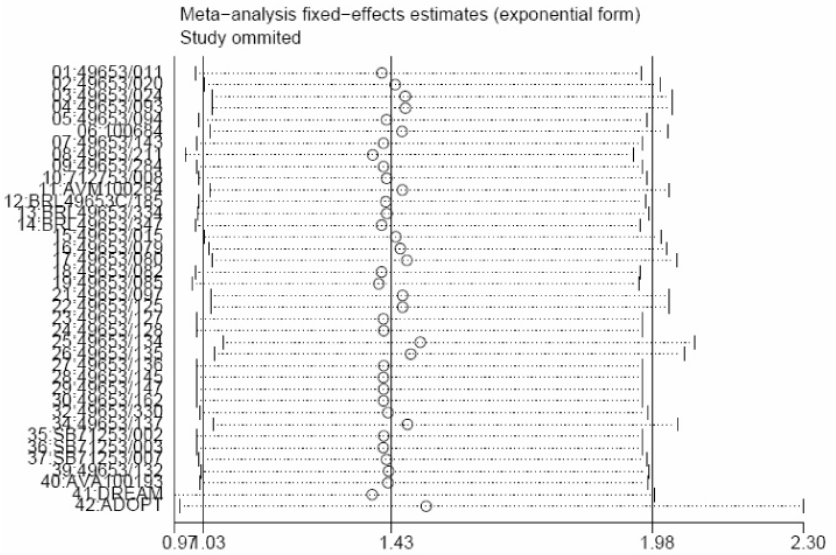

Fig. 5. A plot for finding most influential trial in the result of meta-analysis for the odds ratio (OR) of myocardial infarction (MI) between rosiglitazone group and control group. Each of horizontal line indicates a repeated result of meta-analysis after omitting the corresponding trial in turn. Vertical solid lines represent Peto's pooled estimate and its 95% confidence interval using all of 38 trials.

있으면 이를 메타분석에 포함시켜서 결과의 변화여부를 조사, 가중치의 크기를 바꿔가면서 분석에 사용되는 통계적 모형에 따라 결과가 얼마나 변하는지를 평가[56] 등을 시도해 보기를 추천하고 있다[28].

MI 발생률에 대한 rosiglitazone 안전성 메타분석 결과의 민감도를 평가하기 위해 각 연구를 하나씩 제외시켜 나가면서 메타분석을 반복적으로 수행해보았다(Fig. 5). 즉, Fig. 5의 각 가로줄은 해당 연구가 제외되었을 때의 메타분석 결과를 나타내며, 세로줄인 실선에 해당하는 x-축 수치들은 전체 38개 연구들을 모두 사용한 경우의 Peto 방법에 의한 메타분석 결과(즉, OR 결합추정치 및 95% 신뢰구간, 1.43 (1.03~1.98))를 표시한다. 메타분석 결과는 맨 밑의 가로 줄, 즉, 2개 대규모 임상시험 결과 중 하나인 ADOPT trial이 제외되는 경우에 제일 많이 변화되는 것으로 나타났으며, 이 경우, 즉, ADOPT trial을 제외하고 나머지 37개 연구들만을 사용한 경우의 메타분석 결과는 1.50 (0.98~2.30)으로 통계적인 유의성 여부가 변하였다. 참고로 DREAM 연구를 제외한 경우에는 1.38 (0.97~1.98)이었다.

3. 메타분석 결과의 제시

메타분석 연구를 실시할 때 무엇보다도 중요한 것은 모든 분석과정을 투명하게 제시하고, 연구들의 포함기준 및 제외기준을 명확히 기록해두는 것이다. 또한 이 선택기준들을 적용한 문헌탐색 대상 자료원과 사용된 탐색전략들을 명시해 주어야 하고, 선택된 연구들의 질을 평가하는데 사용된 기준들에 대해서도 자세히 기술해 주어야 한다. 이렇게 하는 이유는 연구자가 논문을 선택하는 과정 중에 본인의 의도와는 상관없이 발생될 수도 있는 bias들을 투명하게 밝혀서 독자들로 하여금 해당 메타분석의 결과에 기초한 근거의 한계들, 이러한 한계들을 극복하기 위한 노력, 그리고 해당 결과로부터 유도된 추론의 신뢰성 등을 평가할 수 있도록 해 주기 위해서이다. 따라서 메타분석 결과를 제시할 때에는 Fig. 4와 같은 그래프들을 적극적으로 활용하는 것이 좋다. 논문을 작성할 때에는 RCT를 대상으로 한 메타분석의 경우, 연구의 질을 향상시키고 그 기준을 제시하기 위해 개발된, QUOROM 지침[57]을, 그리고 관찰연구를 대상으로 한 메타분석의 경우에는 MOOSE 지침[58]을 준수하는 것이 좋다.

4. 메타분석에 대한 비판

메타분석은 현재 가용한 근거들을 결합하는데 매우 효과적으로 사용될 수 있는 방법이긴 하지만, 이에 대한 비판도 있다. 메타분석에 대한 모든 비판들은 근본적으로 자료가 각기 다른 연구들로부터 모아졌을 때에는 통계적인 방법을 실시해도 연구들 간 차이가 완전히 보정되지 못한다는 관점에 기초한다. 이들 비판들 중 가장 주된 것은 메타분석 결과에는 여러 종류의 bias(예를 들면 selection bias, identification bias, language bias, citation bias, publication bias, time-lag bias, database bias, multiple publication bias 등)가 동시에 존재할 수 있다는 점이다[59,60]. 따라서 메타분석에서 얻어지는 결합추정치는 사실상 현실에서는 사용하기에 부적절한, 무의미한 측정값이라는 것이다[25]. 메타분석의 두 번째 문제점으로는 분석에 포함되지 않은 연구들은 처음부터 탐색되지 못한 연구들이기 때문에 이들이 어떤 연구들인지 평가하기가 아예 연구 시작단계부터 불가능하다는 점이다. 사실 전문적인 문헌탐색이 이루어지지 않으면 모든, 적절한, 연구들이 파악되지 못할 수도 있다[44]. 셋째, 메타분석에 포함된 연구들은 모두 자체적으로 내적 타당도(internal validity)와 관련된 문제점들을 가지고 있을 수 있는데, 메타분석 시 이 bias들이 간과될 가능성이 있고, 이는 곧 해당 결과에 영향을 미치게 된다는 점이다[25,44]. 이런 측면 때문에 메타분석 시 연구의 질에 대한 평가가 중요하다. 만일 연구들 간에 이질성의 정도가 너무 심하면 해당 연구결과들을 결합하지 말고 더 많은 연구결과가 얻어질 때까지 해당 처리의 효과에 대한 최종 결정을 유보해야 할 것이다. 넷째, 부적절한 메타분석의 실시 역시 메타분석이 비판을 받게 되는 주된 이유 중 하나이다. 해당 주제에 대한 임상적 지식 및 이론적 통찰력 없이 연구를 수행하는 것, 해당 주제와 관련해 문헌탐색을 적절하게 실시하지 못하는 것, 선정된 논문들로부터 정보를 요약하는 과정을 부주의하게 진행하는 것, 서로 다른 결과변수를 제시한 연구들을 무분별하게 결합하는 것, 중요한 변수들을 미처 고려하지 못하는 것, 메타분석을 수행하는 연구자의 편견이 개입되는 것, 충분한 전문적인 지식 없이 통계적인 방법을 무분별하게 사용하는 것, 그리고 어쩌면 가장 자주 발생하는 문제점인 메타분석 결과의 강점 및 정밀도를 과장되게 서술하는 것 등이 메타분석 결과의 신뢰성을 떨어뜨리는 주요 원인들이라 할 수 있다[44,61]. 사실 메타분석 결과가 잘 수행된 대규모 무작위 임상시험 결과와 다른 경우도 있으며[62], 비슷한 시기에 동일한 문헌탐색을 거친 메타분석 연구들이 서로 모순된 결과를 보이거나 전혀 상반되는 주장을 펼친 경우들이 종종 발견되기도 한다[63~66]. 바로 이러한 불일치들로 인해 메타분석의 신뢰성에 대한 논란이 제기되게 된 것이다.

그러나 메타분석은 주로 아직 논란의 여지가 있는 연구주제나 또는 처리효과의 크기 변화가 심한 연구주제들을 대상으로 수행되는 경우가 많다. 따라서 위와 같은 문제점들은 메타분석 자체의 문제점을 넘어서 연구주제 자체의 한계점인 측면이 있다. 만일 확실한 대규모 임상시험 결과가 존재한다면 굳이 메타분석을 실시할 필요는 없을 것이다. 사실 메타분석과 관련된 논란의 대부분은 잘못 수행된 메타분석 연구에 기인한다고 할 수 있으며, 메타분석 그 자체는 오히려 연구결과들 간의 불일치로 인한 논쟁을 해결해 줄 수 있는 잠재력을 가진 연구방법론으로 간주될 필요가 있다고 본

다. 즉, 연구자들이 메타분석의 연구방향 및 연구방법을 정확히 설정하고, 선택된 각 연구의 특성에 대해 세심한 주의를 기울이며, 사용된 가정들을 명확하게 정리하여, 메타분석 시 발생할 수 있는 bias를 줄이기 위해 노력하고 연구들 간의 이질성을 최소화하면서 연구결과에 대한 해석에 좀 더 세심한 주의를 기울인다면 타당하고 신뢰성 있는 연구결과들이 도출될 가능성은 올라가게 될 것이다. 또한 의학연구 분야에서 메타분석이 활발하게 사용되고 있긴 하지만 이 연구를 한 개인이 수행하는 것은 사실상 쉽지 않다. 따라서 앞에서 언급했듯이 메타분석을 수행하고자 심사숙고 하는 연구자에게는 이 분야에 대해 경험이 있는 동료 임상 연구자들, 문헌탐색 전문가, 그리고 의학통계 전문가와 팀을 이루어 연구를 진행하기를 추천한다.

결 론

메타분석의 결과는 임상적 근거를 설정하는데 중요한 역할을 한다. 비록 방법론의 복잡성이 메타분석의 한계점 중 하나이긴 하지만, 특정 임상적 질문에 대해 명확한 결론을 내려주는 대규모 무작위 임상시험 결과가 없는 한, 메타분석은 해당 질문에 대해 답을 해 줄 수 있는 차선책 중 하나이다. 또한 기존의 표준요법과 비교해 새로운 치료법이 더 효과적인지에 대해 중간 규모의 임상시험 연구들 간에 서로 불일치한 결과를 보이는 경우, 또는 이 연구들에서 보여준 효과 크기들 간에 서로 차이가 나는 경우 등에 있어서 메타분석은 해당 요법이나 중재에 대한 실제 효과를 좀 더 정확히 측정할 수 있게끔 해 줄 것이다. 그러나 체계적 고찰 및 메타분석 결과는 임상적 논리를 보강해주는 데 도움을 주지만 이를 대체할 수는 없다. 따라서 메타분석 결과를 무비판적으로 수용하거나 무감각하게 적용해서는 안 될 것이며, 임상가의 지식, 경험, 가치관 및 임상적 근거 등을 통합한 의사결정이 이루어져야 할 것이다[67]. 메타분석과 관련한 연구들의 수가 급격하게 증가하고 있는 현 시점에서 이제 임상연구자들은 이 방법의 장점 및 제한점에 대해 자세히 알고 있을 필요가 있으며[68], rosiglitazone 메타분석 연구처럼 그 결과가 실제로 임상 진료(clinical practice)에 영향을 미칠 수 있다는 가능성을 감안한다면 메타분석 연구결과를 비판적으로 평가할 수 있는 역량도 키워나가야 할 것이다.

참 고 문 헌

1. Nissen SE, Wolski K: Effect of rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes. NEJM 356:2457-2471, 2007
2. Psaty BM, Furberg CD:. Rosiglitazone and cardiovascular risk. NEJM 356:2522-2524, 2007
3. Psaty BM, Furberg CD: The record on rosiglitazone and the risk of myocardial infarction. NEJM 357:67-69, 2007
4. The RECORD study group: Rosiglitazone evaluated for cardiovascular outcomes-an interim analysis. NEJM 357:28-38, 2007
5. Drazen JM, Morrissey S, Curfman GD: Rosiglitazone -continued uncertainty about safety. NEJM 357:63-64, 2007
6. Nathan DM: Rosiglitazone and cardiotoxicity-weighing the evidence. NEJM 357:64-66, 2007
7. Rosen CJ: The rosiglitazone story - Lessons from an FDA advisory committee meeting. NEJM 357:844-846, 2007
8. Letters to the editor: Rosiglitazone and cardiovascular risk. NEJM 357:930-940, 2007
9. Letters to the editor: Rosiglitazone and the FDA. NEJM 357:1775-1777, 2007
10. The ACCORD study group: Effects of intensive glucose lowering in type 2 diabetes. NEJM 358:2545 -2559, 2008
11. The ADVANCE collaborative group: Intensive blood glucose control and vascular outcomes in patients with type 2 diabetes. NEJM 358:2560-2572, 2008
12. Goldfine AB: Assessing the cardiovascular safety of diabetes therapies. NEJM 359:1092-1095, 2008
13. Pearson K: Report on certain enteric fever inoculation statistics. BMJ 3:1243-1246, 1904
14. Egger M, Davey Smith G: Meta-analysis: potentials and premise. BMJ 315:1371-1374, 1997
15. Lasagna L, Mosteller F, von Felsinger JM, Beecher HK: A study on the placebo response. Am J Med 16:770-779, 1954
16. Glass GV: Primary, secondary and meta-analysis of research. Edu Res 5:3-8, 1976
17. Cochrane AL: Effectiveness and efficiency: Random reflections on the health service. London: Nuffield Provincial Hospitals Trust, 1972
18. Cochrane AL: 1931-1971: A critical review, with particular reference to the medical profession. In: Medicines for the year 2000. London: Office of Health Economics, 1-11, 1979
19. Chalmers I, Sackett D, Silagy C, Maynard A, Chalmers TC: Non-random reflections on health services research. London: BMJ Publishing Group, 1997
20. Sackett DL, Rosenberg WM, Gray JA, Haynes RB,

Richardson WS: Evidence-based medicine: what is it and what it isn't. BMJ 312:71-72, 1996
21. Lee JY: Medical Statistics at a Glance. Seoul, Epublic, 2007
22. Sackett DL, Straus S, Richardson S, Rosenberg W, Haynes RB: Evidence-based medicine: how to practice and teach EBM. 2nd ed. Churchill -Livingston, London, 2000
23. Greenhalgh T: Papers that summarise other papers (systematic reviews and meta-analyses). BMJ 315:672-675, 1997
24. DerSimonian R, Laird NM: Meta-analysis in clinical trials. Control Clin Trials 7:177-186, 1986
25. Egger M, Davey Smith G, Altman D: Systematic reviews in health care: Meta-analysis in context. London, BMJ Publishing Group, 2001
26. L'Abbe KA, Detsky AS, O'Rourke K: Meta-analysis in clinical research. Ann Intern Med 107:224-233, 1987
27. Meade MO: Selecting and appraising studies for a systematic review. Ann Intern Med 127:531-537, 1997
28. Higgins JPT, Green S: Cochrane Handbook for Systematic Reviews of Interventions. version 5.0.0. The Cochrane Collaboration, 2008. Available from http://www.cochrane-handbook.org Accessed Nov 28, 2008
29. Counseil C: Formulating questions and locating primary studies for inclusion in systematic reviews. Ann Intern Med 127:380-387, 1997
30. Egger M, Davey Smith G, Phillips AN: Meta-analysis: principles and procedures. BMJ 315:1533-1537, 1997
31. Simes RJ: Confronting publication bias: a cohort design for meta-analysis. Stat Med 6:11-29, 1987
32. Egger M, Davey Smith G, Schneider M, Minder C: Bias in meta-analysis detected by a simple, graphical test. BMJ 315:629-634, 1997
33. Greenland S: Invited commentary: a critical look at some popular meta-analytic methods. Am J Epidemiol 140:290-296, 1994
34. Schulz KF, Chalmers I, Hayes RJ, Altman DG: Empirical evidence of bias: Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA 273:408-412, 1995
35. Moher D, Pham B, Jones A, Jones A, Cook DJ, Jadad AR, Moher M, Tugwell P, Klassen TP: Does quality of reports of randomized trials affect estimates of intervention efficacy reported in meta-analyses? Lancet 352:609-613, 1998
36. Moher D, Jadad AR, Nichol G, Penman M, Tugwell P, Walsh S: Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists. Control Clin Trials 16:62-73, 1995
37. Juni P, Witschi A, Bloch R, Egger M: The hazards of scoring the quality of clinical trials for meta-analysis. JAMA 282:1054-1060, 1999
38. Moher D, Cook DJ, Jadad AR, Tugwell P, Moher M, Jones A, Pham B, Klassen TP: Assessing the quality of reports of randomised trials: implications for the conduct of meta-analyses. Health Technol Assess 3:1-98, 1999
39. Chalmers TC, Smith H, Blackburn B, Silverman B, Schroeder B, Reitman D, Ambroz A: A method for assessing the quality of a randomized control trial. Control Clin Trials 2:31-49, 1981
40. Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJM, Gavaghan, DJ, McQuay H: Assessing the quality of reports of randomized clinical trials: Is blinding necessary? Control Clin Trials 17:1-12, 1996
41. Stewart LA, Clarke M: Practical methodology of meta-analysis (overviews) using updated individual patient data: Cochrane Working Group. Stat Med 14:2057-2079, 1995
42. Cochrane Collaboration Individual Patient Data Meta -analysis Methods Group. http://www.ctu.mrc.ac.uk/cochrane/ipdmg/ Accessed Nov 28, 2008
43. Hedges LV: Statistical Methodology in Meta-Analysis. Educational Testing Service, Princeton, 1982
44. Sutton AJ, Abrams KR, Jones DR, Sheldon TA, Song F: Methods for meta-analysis in medical research. John Wiley and Sons, New York, 2000
45. Fleiss JL: The statistical basis of meta-analysis. Stat Meth Med Res 2:121-145, 1993
46. Cochran WG: The combination of estimates from different experiments. Biometrics 10:101-129, 1954
47. Higgins JPT, Thompson SG, Deeks J, Altman DG: Statistical heterogeneity in systematic reviews of clinical trials: a critical appraisal of guidelines and practice. J Hlth Serv Res Policy 7:51-61, 2002
48. Higgins JPT, Thompson SG, Deeks J, Altman DG: Measuring inconsistency in meta-analyses. BMJ 327:557-560, 2003

49. Fleiss JL: Analysis of data from multiclinic trials. Control Clin Trials 7:267-275, 1986
50. Furberg CT, Morgan TM: Lessons from overviews of cardiovascular trials. Stat Med 6:295-303, 1987
51. Oxman AD, Guyatt GH: A consumer's guide to subgroup analyses. Ann Intern Med 116:78-84, 1992
52. Schmid CH: Exploring heterogeneity in randomized trials via meta-analysis. Drug Info J 33:211-224, 1999
53. Glasziou PP, Sanders SL: Investigating causes of heterogeneity in systematic reviews. Stat Med 21:1503-1511, 2002
54. Song F: Exploring heterogeneity in meta-analysis: is the L'Abbe plot useful? J Clin Epidemiol 52:725-730, 1999
55. Galbraith RF: A note on the graphical presentation of estimated odds ratio from several clinical trials. Stat Med 7:889-894, 1988
56. Thompson SG: Controversies in meta-analysis: the case of the trials of serum cholesterol reduction. Stat Meth Med Res 2:173-192, 1993
57. The QUOROM group: Improving the quality of reports of meta-analyses of randomized controlled trials: the QUOROM statement. Lancet 354:1896-1900, 1999
58. The MOOSE group: Meta-analysis of observational studies in epidemiology. a proposal for reporting. JAMA 283:2008-2012, 2000
59. Egger M, Davey Smith G: Bias in location and selection of studies. BMJ 316:61-66, 1998
60. Song F, Eastwood AJ, Gilbody S, Duley L, Sutton AJ: Publication and related biases. Health Technol Assess 2004:1-105.
61. Shapiro S: Meta-analysis/Shmeta-analysis. Am J Epidemiol 140:771-778, 1994
62. LeLorier J, Gregoire G, Benhaddad A, Lapierre J, Derderian F: Discrepancies between meta-analyses and subsequent large randomized, controlled trials. NEJM 337:536-542, 1997
63. Rosenfeld RM, Post JC: Meta-analysis of antibiotics for the treatment of otitis media with effusion. Otolaryngol Head Neck Surg 106:378-386, 1992
64. Williams RL, Chalmers TC, Strange KC, Chalmers FT, Bowlin SJ: Use of antibiotics in preventing recurrent acute otitis media and in treating otitis media with effusion: a meta-analytic attempt to resolve the brouhaha. JAMA 270:1344-1351, 1993
65. Kerlikowske K, Grady D, Rubin SM, Sandrick C, Ernster VL: Efficacy of screening mammography: a meta-analysis. JAMA 273:149-154, 1995
66. Smart CR, Hendrick RE, Rutledge JH, Smith RA: Benefit of mammography screening in women ages 40 to 49 years: current evidence from randomized controlled trials. Cancer 75:1619-1626, 1995
67. Cook DJ, Mulrow CD, Haynes RB: Systematic reviews: synthesis of best evidence for clinical decisions. Ann Int Med 126;376-380, 1997
68. Davey Smith G, Egger M: Meta-analysis: unresolved issues and future developments. BMJ 316:221-225, 1998