진단검사의 정확도 평가를 위한 체계적 고찰 방법론

Methodological issues for Systematic Reviews of Diagnostic Tests Accuracy

Abstract

The Cochrane Collaboration says that the Cochrane handbook for diagnostic test accuracy reviews is in development now. That means the methodology of systematic reviews (SR) of diagnostic tests assessments is still a matter of debate. At this point in time, comparison of methodological issues for SR of between interventions and diagnostic tests would be helpful to understand these situations.

Key words: Review literature as topic, Meta-analysis as topic, Diagnostic test, Clinical trial

1. 서론

최근 들어 비교효과연구 (Comparative Effectiveness Research, CER)와 의료기술평가 연구 (Health Technology Assessment, HTA)가 활성화 되면서 그 연구방법론으로 메타분석 (meta-analysis)을 적용한 체계적 고찰 연구 (systematic review)가 크게 부각되고 있다 [1-3]. 특히 약물이나 시술 같은 개입(intervention)의 효과를 비교하는 무작위 배정 임상 시험을 집중적으로 다루는 방법론에 있어, 코크란 연합 (Cochrane Collaboration)의 Higgins & Green 이 저자인 개입연구 (intervention)를 대상으로 한 체계적 고찰의 지침서 (Handbook)가 개발되어 확산된 것이 주된 배경이라 본다 [4].

그런데 CER이나 HTA 연구에는 개입연구뿐만 아니라, 진단검사도 분석 대상으로 삼고 있다. 사실 현대의학은 진단학 (Diagnostics) 이라 할 정도로 진단이 정확해야만 제대로 된 치료서비스를 제공할 수 있고, 생존율 등에서 최선의 성과를 얻어낼 수 있기 때문이다. 따라서 진단검사의 평가 (Diagnostic Test Assessment, DTA)에 대한 체계적 고찰 연구방법론도 당연히 필요하다. 그런데 DTA 에 대한 코크란 연합의 홈페이지에서 알 수 있듯이 진단검사와 관련한 방법론은 현재 개발 중에 있다 [5,6]. 이런 현 시점에서 코크란 연합이 지금까지 제안한 DTA 관련 개념들과 방법들을, 개입연구의 방법론과 상호 대조하여 살펴보고자 한다.

2. 체계적 고찰 진행 단계별 방법론 검토

Table 1은 체계적 고찰의 수행 단계에 따라 관련한 개념과 지표들을 개입연구와 진단검사로 나누어 상호 대조시켜 정리한 표이다. 이 표 내용을 중심으로 관련 내용들을 펼치

고자 한다.

가. 질문 설정

체계적 고찰의 첫 시작은 답할 수 있는 질문들 (Answerable questions)로 전환시키는 것이다. 이에 개입연구는 환자 특성 (Patient), 개입 처치 (Intervention), 대조 처치 (Comparator), 예상하는 성과 (Outcomes)의 4가지 관련어 앞 자를 따서 'PICO' 란 도구를 제시하고 있다 [7].

반면 진단검사는 'PPP-IP-PTR' 의 8가지 관련 내용을 제시토록 요구하고 있다 [6]. 첫 번째 P는 환자 특성으로 개입연구의 P와 동일하지만, 두 번째와 세 번째 P는 환자들의 주된 증상이나 증후를 제시하는 Presentation, 과 해당 환자를 진단할 때 사용한 Prior tests를 각각 의미한다. 4번째 I는 체계적 고찰을 하려는 검사 (Index test)이며, 5번 째 C는 Comparator test로 통상적으로 시행하고 있어 Index test와 비교하려는 검사이다. 따라서 진단검사의 IC는 개입연구의 IC와 짝을 지을 수 있겠다. 6번째 P는 연구목적 Purpose이며, 3 가지로 대분할 수 있다. ① 기존의 Comparator test를 Index test로 대치하는 것 (Replacement) ② Index test를 시행하여 양성이 나온 대상에게 Comparator test를 시행하여 보다 세분된 진단을 얻으려는 것 (Triage) ③ Comparator test를 시행하여 음성이 나온 대상에게 Index test를 시행하여 위음성을 낮추려는 것 (Add-on). 7번째 T는 target disorder는 새로운 검사로 진단하려는 특정 질환을 뜻하여 개입연구 PICO의 O와 개념상 짝지어 볼 수 있다. 마지막 8번째 R은 Reference standard로 확진검사 (Gold standard)를 의미한다.

이처럼 진단검사의 평가를 위해 체계적 고찰을 하려면 검토해야 할 내용들이 매우 다양하다는 것을 짐작할 수 있다. 특히 검사에 있어 Prior test, Index test, Comparator test, Reference standard 같이 4가지 종류를 정리하도록 요구하고 있어, 이에 대한 개념 구분이

필요하다. 예를 들어 유방촬영술상 치밀유방으로 나온 고위험 유방암 검진자들에게 추가 검사로 유방초음파를 하는 경우와 유방MRI를 할 경우 유방암 진단의 타당성을 평가하는 연구를 수행한다고 가정한다면, Index text는 유방MRI, Comparator test는 유방초음파, Prior test는 유방촬영술, Reference standard는 유방조직의 해부병리 판독이 될 것이다.

나. 논문 검색

검색 전략에 사용할 주요 검색어로는 개입연구라면 PICO의 I에 해당하는 개입이 될 것이지만, 진단검사에서는 index test (I) 와 target disorder (T)가 해당될 것이다. 또한 개입연구는 대부분 무작위배정 임상시험 (Randomized Controlled Trial, RCT) 연구 설계를 적용하고 있기 때문에 연구설계 방법을 필터링 하면서 개입에 관한 중심어로 검색하는 것이 효과적인 검색 전략이 된다. 그러나 진단검사는 비교를 위한 RCT뿐만 아니라 단면연구 (cross-sectional study) 등의 다양한 연구 설계를 적용하기 때문에, 연구설계를 필터링하는 검색 전략은 무의미하다.

다. 개별논문의 평가 및 정보 추출

개별연구의 질적 수준을 평가하는 도구로 개입연구는 Cochrane 연합이 제시한 ROB (Risk of Bias가 있다면 [8], 진단검사는 QUADAS-2 (the Quality Assessment of Diagnostic Accuracy Studies) 도구가 개발되어 있다 [9]. 2003년도에 개발한 QUADAS를 수정 보완하여 2011년도에 발표한 QUADAS-2은 대상자 선정 (patient selection), index test, reference standard, 연구 수행과정 (flow and timing)의 4가지 영역으로 나누고 있으며, 이중 앞의 3가지 영역에 속한 질문들에 대하여 Yes (High), No (Low), Unclear의 3가지 중 하나를 답변토록 요구한다 [10]. QUADAS-2를 국내 연구진들이 이해하기 쉽고 활용하기 좋게 번안하는 작업이 있기를 기대한다.

선정된 관련 논문들의 결과에서 얻어낼 정보로는, 개입연구일 경우 처치군 (treatment group)과 비교군 (control group)간의 반응 분율 (%)에 관련한 수치들이다. 반면 진단검사일 경우 민감도 (sensitivity)와 특이도 (specificity)가 된다 [11]. 진단검사에서는 예측도 (predictive value) 결과도 있지만 연구 대상자의 유병률 (prevalence)에 따라 변하는 값이기에 체계적 고찰의 특성에 적합하지 않다 [12]. 반면 민감도와 특이도는 유병률과 무관하기에 우선적으로 활용을 하게 된 것이다 [13]. 그렇지만 이들 또한 기준점 (threshold)에 따라 변동을 한다는 한계를 가지고 있어, Receiver operator characteristic curve (ROC 곡선)를 같이 제시하는 것이 필수적이다 [14].

추출한 정보로부터 새로운 의미들을 알아보기 위하여 관련 지표들을 산출하는데, 개입연구는 처치군과 비교군의 반응률의 차이의 역수를 구하여 Number Needed to Treat (NNT)로 제시한다 [15]. 반면 진단검사는 True results에 속하는 민감도와 특이도의 곱을 False results에 속하는 수치들의 곱으로 나누어 Diagnostic Odds Ratio (DOR)이란 지표를 따로 산출한다 [16,17]. 이 값은 2*2 표에서 얻어내는 ad/bc와 같은 수식형태를 갖기에 OR이라 하며, 이 값이 클수록 민감도와 특이도가 상대적으로 더 크다는 뜻이다. 달리 해석하자면 ROC 곡선에서 좌상 (Left & Upper)의 꼭지점으로 더 접근한다는 것을 의미하며, 그만큼 곡선아래의 면적 (Area under the curve, AUC)이 커진다는 의미한다 [14].

추출한 정보들을 일목요연하게 보여주기 위하여 개입연구는 Forest plot을 사용한다 [18]. 그런데 진단검사는 민감도와 특이도라는 두 가지 정보를 같이 보여주는 Coupled forest plot으로 제시한다 [19]. 또한 앞서 언급한 것처럼 민감도와 특이도는 기준점에 따라 변동하기 때문에 summary ROC (SROC) 곡선을 같이 제시한다 [20]. 논문 대상자 수나 표준오차에 따라 표기되는 기호의 크기를 달리 할 수 있다.

라. 메타분석

메타분석을 하려면 대상 논문들 간의 이질성을 반드시 확인해야 한다. 개입연구는 최근 I^2 통계값을 활용하여 그 정도를 평가하고 있다 [21]. 이에 맞추어 동질성이 확보되면 fixed effect model에 따라, 이질성이 확인되면 random effect model에 따라 요약 통계값을 산출하는 것이다.

그렇지만 진단검사는 민감도와 특이도의 trade-off 속성 등의 한계를 감안하여 특별한 경우가 아니면 이질성이 있다고 간주하고 있다. 특히 고혈압 진단기준처럼 기준점이 계속 달라져 온 경우에는 이를 반영하는 공변수 (covariate)에 따라 하부군 분석 (subgroup analysis)를 해야만 한다 [10]. 따라서 이질성을 평가하는 통계법이 아직 정해진 것이 없으며, 다층모델 (Hierarchal random effect model)에 따른 추가 분석을 대부분 요구하고 있다. 현재 Bivariate method와 Rutter & Gatsonis HSROC method 두 가지 방식이 개발되어 있는데, 실무 적용에 있어 이 둘 간의 차이는 산출에 사용하는 통계값이 다른 것이다 [6]. Bivariate method는 민감도와 특이도를 그대로, HSROC method 는 threshold와 DOR 를 사용한다 [20]. 그런데 RevMan 5.3에서는 이 두 가지 분석 모두를 직접 지원하지 않으며, SAS (PROC NLMIXED) 나 STATA (METANDI)에서 분석하여 얻어낸 통계값을 추가로 입력하면 RevMan은 그 요약통계값을 보여주고 있는 수준이다 [21]. 만약에 대상 논문수가 적고 기준점의 변동이 없다는 전제라면 Moses-Littenberg SROC를 요약 통계값으로 활용할 수는 있겠다.

마. 결과작성

체계적 고찰의 대상이 되는 원저 (original article)에 있어 개입연구의 결과를 제시하는 지침으로 CONSORT (Consolidated Standards of Reporting Trials) 가 있는 반면 [23], 진단검사는 STARD (the Standards for Reporting of Diagnostic Accuracy)가 있다 [24]. 그리고 개입연구

를 대상으로 체계적 고찰을 시행하여 얻어낸 결과를 보고하는 지침으로 PRISMA (Preferred Reporting Items for Systematic reviews and Meta-analysis)가 개발된 반면 [25], 진단검사에 관한 체계적 고찰 보고 지침은 아직 없다. 그리고 출판과정에서 생길 수 있는 오류를 간접 확인하는 방법으로 개입연구는 Funnel Plot을 활용할 수 있으나, 진단검사는 아직 이에 대한 평가를 할 수 있는 도구가 개발된 것이 없다.

3. 결론 및 제언

진단검사 논문들의 체계적 고찰을 위한 연구방법론은 현재 개발 중이라는 것은 그만큼 검토해야 할 것이 많다는 것이다. 관련 전문가들 간의 이견을 좁히지 못하고 있을 뿐만 아니라, 진단검사가 갖는 특별한 속성 때문에 개입연구보다 더 극복해야 할 방법론 이슈들이 아직 산재해 있기 때문이다 [26]. 이번 고찰에서 제시된 내용들이 이후에 얼마든지 바뀔 수 있다는 가능성은 분명 열려있다. 그럼에도 불구하고, 현 시점에서 이렇게 방법론을 비교 고찰해 본 것은 국내 연구진들이 이에 대한 관심을 가지고 적극 개입하기를 바라는 의도이다. 통계학 전공자뿐만 아니라 역학자들도 진단검사의 체계적 고찰을 많이 시도할 자극이 되기를 바라며 이만 줄인다.

감사의 글

이 논문은 2014학년도 제주대학교 학술진흥연구비 지원사업에 의하여 연구되었음

References

1. Drummond MF, Schwartz JS, Jőnsson B, Luce BR, Neumann PJ, Siebert U, et al. Key principles for the improved conduct of health technology assessments for resource allocation decisions. Int J Tech Assess Health Care 2008;24:244-258.
2. Manchikanti L. Evidence-based medicine, systematic reviews, and guidelines in interventional pain management, part I: introduction and general considerations. Pain Physician 2008;11:161-186.
3. Kim SY, Park JE, Seo HJ, Seo HS, Son HJ, Shin CM, et al. Development of Manual for Systematic reviews and clinical practice guideline; 2010 [cited 2014 Aug 11]. Available from: http://www.neca.re.kr/center/researcher/report_view.jsp?boardNo=GA&seq=17&q=626f6172644e6f3d4741.
4. Higgins JPT, Green S. Cochrane handbook for systematic reviews of interventions. The Cochrane Collaboration. John Wiley & Sons: Chichester, UK; 2008.
5. The Cochrane Collaboration. Cochrane handbook for diagnostic test accuracy reviews. [cited 2014 Aug 11]. Available from: http://www.cochrane.org/editorial-and-publishing-policy-resource/cochrane-handbook-diagnostic-test-accuracy-reviews.
6. Diagnostic Test Accuracy Working Group. Handbook for DTA reviews. [cited 2014 Aug 11]. Available from: http://srdta.cochrane.org/handbook-dta-reviews.
7. Tseng TY, Dahm P, Poolman RW, Preminger GM, Canales BJ, Montori VM. How to use a systematic literature review and meta-analysis. J Urol 2008;180:1249-1256.
8. Higgins JPT, Altman DG, Gøtzsche PC, Jüni P, Moher D, Oxman AD, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ 2011;343:d5928.
9. Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reltsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011;155:529-536.
10. Schuetz GM, Zacharopoulou NM, Schlattmann P, Dewey M. Meta-analysis: noninvasive coronary angiography using computed tomography versus magnetic resonance imaging. Ann Intern Med

2010;152:167-177.

11. Honest H, Khan KS. Reporting of measures of accuracy in systematic reviews of diagnostic literature. BMC Health Services Research 2002;2;4.
12. Montori VM, Wyer P, Newman TB, Keitz S, Guyatt G, for the Evidence-Based Medicine Teaching Tips Working Group. Tips for learners of evidence-based medicine: 5. The effect of spectrum of disease on the performance of diagnostic tests. CMAJ 2005;173:385-390.
13. Schünemann HJ, Oxman AD, Brozek J, Glasziou P, Jaeschke R, Vist GE, et al. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. BMJ 2008l;336:1106-1110.
14. Deeks JJ. Systematic reviews of evaluations of diagnostic and screening tests. BMJ 2001;323:157-162.
15. Cook RJ, Sackett DL. The number needed to treat: a clinically useful measure of treatment effect. BMJ 1995;310:452-454.
16. Glas AS, Lijmer JG, Prins MH, Bonsel GJ, Bossuyt PMM. The diagnostic odds ratio: a single indicator of test performance. J Clin Epidemiol 2003;56:1129-1135.
17. Devillé WL, Buntinx F, Bouter LM, Montori VM, de Vet HC, van der Windt DA, et al. Conducting systematic reviews of diagnostic studies: didactic guidelines. BMC Med Res Methodol 2002;2:9.
18. Engberg S. Systematic reviews and meta-analysis. J Wound Ostomy Continence Nurs 2008;35:258-265.
19. Leeflang MM, Debets-Ossenkipp YJ, Visser CE, Scholten RJ, Hooft L, Bijlmer HA, et al. Galactomannan detection for invasive aspergillosis in immunocompromized patients. Cochrane Database Syst Rev 2008;4:CD007394.
20. Irwig L, Tosteson ANA, Gatsonis C, Lau S, Colditz G, Chalmers TC, et al. Guidelines for meta-analyses evaluating diagnostic tests. Ann Intern Med 1994;120:667-676.
21. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med 2002;21;1539-

1558.

22. Zhang Z, Lu B, Sheng X, Jin N. Accuracy of stroke volume variation in predicting fluid responsiveness: a systematic review and meta-analysis. J Anesth 2011;25:904-916.
23. Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elboume D, et al. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. Ann Intern Med 2001;134:663-694.
24. Simel DL, Rennie D, Bossuyt PM. The STARD statement for reporting diagnostic accuracy studies: application to the history and physical examination. J Gen Intern Med 2008;23:768-774.
25. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JPA, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. Ann Intern Med 2009;151:W65-W94.
26. Oakley A, Strange V, Bonell C, Allen E, Stephenson J, RIPPLE Study Team. Process evaluation in randomised controlled trials of complex interventions. BMJ 2006;332:413-416.

Table 1. Comparison of issues related to systematic reviews for intervention trials and diagnostic test studies

STEP	Issues	for Intervention	for Diagnostic test
Ask			
	Making Questions	PICO	PPP-IP-PTR
Acquire			
	Main keyword	Intervention	Index test & Target disorder
	Searching	Filtering	No filtering
Assess			
	Quality Level	ROB	QUADAS-2
	Extracting Results	Proportion of Response (%)	Sensitivity & Specificity
	New Index	NNT	DOR
	Summary Figures	Forest Plot	Coupled Forest Plot & SROC
Analysis			
	Heterogeneity index	I^2	(SROCs by prediction region)
	on Homogeneous	Fixed effect model	(Moses-Littenberg SROC)
	on Heterogeneous	Random effect model	Hierarchal models
Report			
	Standard for original article	CONSORT	STARD
	Standard for summary results	PRISMA	Not available
	Publication bias	Funnel Plot	Not available