epiH
Epidemiology and Health

Epidemiology and Health
Volume: 36, Article ID: e2014016, 4 pages
http://dx.doi.org/10.4178/epih/e2014016

METHODS

Open Access

An overview of systematic reviews of diagnostic tests accuracy

Jong-Myon Bae

Department of Preventive Medicine, Jeju National University School of Medicine, Jeju, Korea

The Cochrane Collaboration says that the Cochrane handbook for diagnostic test accuracy reviews (DTAR) is currently in development as per the Cochrane Collaboration. This implies that the methodology of systematic reviews (SR) of diagnostic test accuracy is still a matter of debate. At this point, comparison of methodologies for SR in case of interventions as against diagnostics would be helpful to understand DTAR.

KEY WORDS: Review literature as topic, Meta-analysis as topic, Diagnostic test, Clinical trial

INTRODUCTION

Recently, as comparative-effectiveness research (CER) and health-technology assessment (HTA) are being widely implemented, the need for systematic reviews (SR) with meta-analysis is growing [1-3]. In particular, SR methodology for randomized controlled clinical trials (RCT), which compare effectiveness of drug or procedural interventions, has been established on the Cochrane Handbook for Systematic Reviews of Interventions by Higgins et al. [4].

However, CER and HTA involve analyses of diagnostic tests accuracy as well as that of interventional trials. In reality, modern medicine includes a majority of diagnostics: one can only administer right treatment and obtain positive results—in regards to survival rates, for instance—with the help of accurate diagnosis. Accordingly, SR methodology regarding diagnostic test assessments (DTA) is clearly required. However, SR methodology of DTA is currently in development, as indicated on the Cochran Collaboration website [5,6]. At such a point, this article will overview DTA-related concepts and issues of SR methodology proposed by the Cochran Collaboration.

Correspondence: Jong-Myon Bae
Department of Preventive Medicine, Jeju National University School of Medicine, 102 Jejudaehak-ro, Jeju 690-756, Korea
Tel: +82-64-755-5567, Fax: +82-64-725-2593, E-mail: jmbae@jejunu.ac.kr

Received: Aug 13, 2014, Accepted: Aug 29, 2014, Published: Aug 29, 2014
This article is available from: http://e-epih.org/

STEPS OF SYSTEMATIC REVIEWS

Table 1 shows a comparison of the concepts and indicators detailing each procedural step in conducting SR for effectiveness of interventional trials and accuracy of diagnostic tests. Based on this table, the following content may be proposed.

Making the answerable questions

The first step in a SR is to convert facing problems into some answerable questions. The patient or population, intervention, comparator, outcomes (PICO) method is being postulated as a viable tool for this process in SR of interventional trials [7].

The SR of DTA, however, begin with addressing the 8 aspects of 'PPP-ICP-TR' [6]. The first 'P' refers to patient characteristics, thus coinciding with the 'P' of the PICO, but the second and third 'P' refer to presentation, which concerns a patient's major symptoms, as well as prior tests, which are used for patient diagnosis. The 'I' refers to index tests, which will be used in conducting systematic reviews, while the 'C' refers to comparator tests, which are regular procedures comparative to the index test. Accordingly, the 'IC' for DTA may correspond to the 'IC' for interventional trials. The last 'P' stands for 'purpose', which may be divided largely into 3 parts: 1) changing the conventional comparator test into the index test (replacement), 2) conducting comparator tests on those who tested positively in index tests, as to obtain a more accurate diagnosis (triage), and 3) conducting index tests on those who tested negatively in com-

Table 1. Comparison of issues related to systematic reviews (SR) of intervention trials and diagnostic tests

Step	Issues	SR of intervention trials	SR of diagnostic tests
Ask	Making questions	PICO	PPP-ICP-TR
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	hierarchical models
Report	Standard for original article	CONSORT	STARD
	Standard for summary results	PRISMA	Not available
	Publication bias	Funnel plot	Not available

ROB, risk of bias; QUADAS-2, Quality Assessment of Diagnostic Accuracy Studies-2; NNT, number needed to treatment; DOR, diagnostic odds ratio; SROC, summary receiver operator characteristic curve; CONSORT, Consolidated Standards of Reporting Trials; STARD, Standards for Reporting of Diagnostic Accuracy; PRISMA, Preferred Reporting Items for Systematic reviews and Meta-analysis.

parator tests, as to reduce false-negative results (add-on). The 'T' stands for the 'target disorder' of any given SR and corresponds with the 'O' of the PICO method. The final 'R' refers to the 'reference standard', or, more specifically, the gold standard.

Indeed, a highly diverse range of information must be examined in order to conduct SR of DTA. Notably, addressing the 4 categories of test–namely, the prior test, index test, comparator test, and reference standard–requires that the concepts be clearly differentiated according to context. For instance, when conducting SR for choice between breast ultrasonography and breast magnetic resonance imaging (MRI) as additional examinations for diagnosis of breast cancer on women who showed dense mammography, the index test, comparator test, prior test, and reference standard would belong to breast MRI, breast ultrasonography, mammography, and pathologic results of breast tissues, respectively.

Searching literature

While key words for performing SR of interventional trials might include 'intervention' (the 'I' of the 'PICO' method), while those for diagnostic studies might include 'index test' (the 'I' of 'PPP-ICP-TR') and the target disorder (the 'T' of 'PPP-ICP-TR'). Moreover, in interventional trials, filtering study design for RCT while focusing on the topics concerning the intervention can be an effective search strategy, as most interventional trials use the RCT design. However, as diagnostic tests utilize a diverse range of research design, such as cross-sectional studies (and not just comparative RCTs), it is meaningless to filter for research designs when searching for literature concerning diagnostic tests.

Evaluating individual article and extracting information

As for tools that evaluate the quality of each article, risk of bias (ROB) if applicable, as proposed by Higgins et al. [8], as well as Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) [9], have been developed for interventional trials and diagnostic tests, respectively. QUADAS-2 is the revised, 2011 version of the 2003 QUADAS and consists of 4 dimensions – patient selection, index test, reference standard, and finally, flow and timing, the first 3 of which requires an answer among the 3 available responses (yes/high, no/low, and unclear) [10]. An accessible Korean adaptation of the QUADAS-2 would be both, convenient and highly beneficial.

As for data extraction in SR of interventional trials, response rates (%) among the treatment and control groups should be obtained from the articles chosen for review. With respect to diagnostic tests, however, sensitivity and specificity of the tests are essential [11]. Diagnostic tests also provide predictive values, but these change according to disease prevalence and thus, are not appropriate for use in SR of DTA [12]. Likewise, sensitivity and specificity have been used, precisely as they are not associated with disease prevalence [13]. However, these factors also change according to threshold level and therefore, receiver operator characteristic (ROC) curves should necessarily accompany their use [14].

When calculating relevant indices in order to detect new information in the extracted data, the number needed to treat in interventional trials should be calculated from the reciprocal values corresponding to the difference in the response rates of the treatment and control groups [15]. In contrast, SR of DTA could rather calculate a diagnostic odds ratio (DOR) by dividing the product of sensitivity and specificity (true results) by the product of the values that count as false results [16,17]. This value is referred to as an OR, as it is in the same form as ad/bc in a 2×2 table: the larger the value, the higher will sensitivity and specificity be in relation to each another. In other words, the larger the value, the closer one will approach the left upper

maximum of the ROC curve and consequently, the larger the area under the curve will be [14].

In order to clearly display the extracted data, SR of interventional trials employ forest plots [18]. However, SR of DTA use coupled forest plots to show information concerning both sensitivity and specificity [19]. In addition, because sensitivity and specificity singularly change according to threshold level, summary ROC (SROC) curves accompany the plots [20]. The size of the mark may be changed according to the sample size of articles selected or standard error.

Meta-analysis

In order to conduct a meta-analysis, heterogeneity among the selected articles must be examined. Currently, SR of interventional trials assess heterogeneity using I^2 values [21]. Summary statistics may be calculated according to fixed-effect models, if homogeneity is confirmed, or according to random-effect models, if heterogeneity is confirmed.

However, taking the trade-offs concerning sensitivity and specificity into the account, SR of DTA assume heterogeneity, except in special cases. In particular, when thresholds, such as standards in hypertension diagnoses, continuously change over time, a subgroup analysis must be conducted according to the covariate that reflects this change [10]. Thus there are no statistical methods designated to assess heterogeneity in SR of DTA, and additional analyses involving hierarchical random-effect models are required mostly. Currently, the two methods such as bivariate method and Rutter & Gatsonis HSROC method have been developed for this purpose. They use different statistical values for calculation [6]. The bivariate method uses sensitivity and specificity, while the HSROC method uses thresholds and DOR [20]. Though, RevMan 5.3 supports neither method of analysis directly, when the statistical estimates from SAS PROC NLMIXED (SAS Inc., Cary, NC, USA) or STATA METANDI (StataCorp, College Station, TX, USA) are additionally entered, RevMan can calculate summary statistics [21]. If fewer studies and fixed threshold were used, the Moses-Littenberg SROC might be useful as well, for summary statistics.

Reporting

Reporting results of interventional trials and diagnostic tests may follow the guidelines postulated by Consolidated Standards of Reporting Trials (CONSORT) [23] and Standards for Reporting of Diagnostic Accuracy (STARD) guidelines [24], respectively. Additionally, while Preferred Reporting Items for Systematic reviews and Meta-analysis (PRISMA) is a guideline for reporting results of SR of interventional trials [25], there is yet none available for that of DTA. Moreover, while funnel plots may be used to indirectly check for publication bias in SR of interventional trials, there is no tool currently available for such evaluations in SR of DTA.

CONCLUSIONS AND SUGGESTIONS

The fact that methodology regarding SR of DTA is still in development implies that several issues remain unaddressed. Experts are yet to reach a consensus and the complex nature of DTA increase the quantum of issues than those in the case of interventional trials [26]. Furthermore, the content discussed in the present study opens for changes at any time. Nevertheless, I deliberated on the methodology for SR of DTA so as to encourage Korean researchers to take interest and actively participate in the process of refining this methodology. Finally, we hope that epidemiologists and biostatisticians will attempt several SR in near future.

ACKNOWLEDGEMENTS

This study was funded by the Jeju National University academic research support program.

CONFLICT OF INTEREST

The author has no conflicts of interest to declare for this study.

SUPPLEMENTARY MATERIAL

Supplementary material is available at http://www.e-epih.org/.

REFERENCES

1. Drummond MF, Schwartz JS, Jönsson 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 Technol 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. National Evidence-based Healthcare Collaborating Agency. 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 (Korean).
4. Higgins JP, Green S; Cochrane Collaboration. Cochrane handbook for systematic reviews of interventions. Chichester: Wiley-Blackwell; 2008.
5. Cochrane Collaboration. Cochrane handbook for diagnostic test accuracy reviews [cited 2014 Aug 11]. Available from: http://www.co-

chrane.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 JP, 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 AW, Westwood ME, Mallett S, Deeks JJ, Reitsma 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 Serv Res 2002;2:4.
12. Montori VM, Wyer P, Newman TB, Keitz S, Guyatt G; 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 2008;336:1106-1110.
14. Deeks JJ. Systematic reviews in health care: 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 PM. 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: studies of studies. J Wound Ostomy Continence Nurs 2008;35:258-265.
19. Leeflang MM, Debets-Ossenkopp 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 AN, Gatsonis C, Lau J, Colditz G, Chalmers TC, et al. Guidelines for meta-analyses evaluating diagnostic tests. Ann Intern Med 1994;120:667-676.
21. Higgins JP, 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, Elbourne 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 JP, 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.