ELSEVIER

CrossMark

Journal of Clinical Epidemiology 67 (2014) 1200–1209

Journal of Clinical Epidemiology

Systematic overview finds variation in approaches to investigating and reporting on sources of heterogeneity in systematic reviews of diagnostic studies

Christiana A. Naaktgeboren[a,*], Wynanda A. van Enst[b,c], Eleanor A. Ochodo[c,d], Joris A.H. de Groot[a], Lotty Hooft[b,c], Mariska M. Leeflang[b,c], Patrick M. Bossuyt[c], Karel G.M. Moons[a], Johannes B. Reitsma[a]

[a]*Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Universiteitsweg 100, 3584 CG Utrecht, The Netherlands*
[b]*Dutch Cochrane Centre, University Medical Center Utrecht, Universiteitsweg 100, 3584 CG Utrecht, The Netherlands*
[c]*Department of Clinical Epidemiology, Biostatistics and Bioinformatics (KEBB), Academic Medical Centre (AMC), University of Amsterdam (UvA), Meibergdreef 9, 1105 AZ Amsterdam, The Netherlands*
[d]*Department of Interdisciplinary Health, Faculty of Medicine and Health Sciences, Stellenbosch University, PO Box 19063, Francie van Zijl Drive, TYGERBERG 7505, South Africa*

Accepted 12 May 2014; Published online 22 July 2014

Abstract

Objectives: To examine how authors explore and report on sources of heterogeneity in systematic reviews of diagnostic accuracy studies.

Study Design and Setting: A cohort of systematic reviews of diagnostic tests was systematically identified. Data were extracted on whether an exploration of the sources of heterogeneity was undertaken, how this was done, the number and type of potential sources explored, and how results and conclusions were reported.

Results: Of the 65 systematic reviews, 12 did not perform a meta-analysis and eight of these gave heterogeneity between studies as a reason. Of the 53 reviews containing a meta-analysis, 40 explored potential sources of heterogeneity in a formal manner and 27 identified at least one source of heterogeneity. The reviews not investigating heterogeneity were smaller than those that did (median [interquartile range {IQR}], 8 [5–15] vs. 14 [11–19] primary studies). Twelve reviews performed a sensitivity analysis, 25 stratified analyses, and 19 meta-regression. Many sources of heterogeneity were explored compared with the number of primary studies in a meta-analysis (median ratio, 1:5). Review authors placed importance on the exploration of sources of heterogeneity; 37 mentioned the exploration or the findings thereof in the abstract or conclusion of the main text.results

Conclusion: Methods for investigating sources of heterogeneity varied widely between reviews. Based on our findings of the review, we made suggestions on what to consider and report on when exploring sources of heterogeneity in systematic reviews of diagnostic studies.

Keywords: Meta-analysis; Diagnostic techniques and procedures/standards; Sensitivity and specificity; Data interpretation; Statistical; Bias (epidemiology)

1. Introduction

As with any review, the results between studies in a review of diagnostic tests are likely to be different, also referred to as variability or heterogeneity in results. Some heterogeneity in the results between studies can be expected simply because of chance variation. Even if studies are methodologically identical and carried out in the same population, their results will vary because each study only observes a finite sample from the total population of interest. This variation is known as chance variation and is directly linked to the sample size of a study.

Statistical tests and measurements (such as Cochran Q test or I^2) are often used to conclude whether there is more heterogeneity than is expected because of chance alone [1]. If there is more heterogeneity than expected because of chance alone, this is termed systematic differences,

Funding: The Dutch Organization for Scientific Research (De Nederlandse Organisatie voor Wetenschappelijk Onderzoek), projects 9120.8004 and 918.10.615.

Conflict of interest: None of the authors have any potential conflicts of interest to declare.

* Corresponding author. Tel.: +31-88-75-681-81; fax: +31-88-75-680-99.
E-mail address: c.naaktgeboren@umcutrecht.nl (C.A. Naaktgeboren).

0895-4356/$ - see front matter
http://dx.doi.org/10.1016/j.jclinepi.2014.05.018

What is new?

Key findings

- A wide variety of approaches are used for exploring sources of heterogeneity in reviews of diagnostic studies.
- Exploring and reporting sources of heterogeneity is complex in reviews of diagnostic tests as they typically focus on two potential correlated outcomes (sensitivity and specificity).

What this adds to what was known?

- Inspired by the variety of approaches observed in this review, a list of items to consider and report on when exploring sources of heterogeneity in diagnostic reviews was developed.

What is the implication and what should change now?

- Further guidance for exploring sources of heterogeneity could improve the strength and usefulness of systematic reviews of diagnostic studies.

statistical heterogeneity, or "true" heterogeneity. In a random-effects model, this "true" heterogeneity is anticipated, and such models then estimate its magnitude with a metric known as τ^2 or the between-study variance [2].

When there are indications that there is "true" heterogeneity, it is likely that something is causing the heterogeneity (eg, the index test's performance varies between settings or the study designs differ between studies), and reviewers are encouraged to look into the possible causes of this heterogeneity [3,4]. Unexplained heterogeneity in a review usually results in a downgrading of the quality of the evidence it provides [5,6]. Identifying whether there are systematic differences in accuracy of the index test between studies is an important step in translating the results of the review to clinical practice.

The pooling of the results from diagnostic studies in a meta-analyses has an additional level of complexity compared with intervention meta-analyses in that there are usually two correlated analytical outcome measures of interest, namely the sensitivity and specificity of the index test. Similar to intervention reviews, there can be many causes for true heterogeneity, including both clinical and nonclinical factors, such as age, disease spectrum, or study design characteristics. However, a special additional source of heterogeneity that reviews of diagnostic studies may present is related to having two correlated outcomes of interest. Sensitivity and specificity are often negatively correlated because of implicit or explicit differences in the index test threshold. This so-called threshold effect adds an additional layer of complexity to the exploration of sources of heterogeneity in meta-analyses of diagnostic studies.

Although guidelines for investigating the sources of heterogeneity in results in systematic reviews of interventions have been established [7], this is not yet the case for systematic reviews of diagnostic studies. The number of systematic reviews of diagnostic studies published each year is rapidly increasing, and the methods for performing such studies have seen many technical developments over the past years [3,8].

The aim of this methodological review of the literature was to document how sources of heterogeneity are currently being explored in systematic reviews of diagnostic accuracy studies and to propose a list of items for researchers to consider and report on when performing such an exploration.

2. Methods

2.1. Overarching project

This study was a part of a metaepidemiologic project on systematic reviews of diagnostic studies. The goal of this project was to investigate several methodological topics such as small sample size effects, time lag bias, quality assessment, and how to interpret tests and measurements of heterogeneity.

2.2. Selection of review articles

Systematic reviews of diagnostic test accuracy studies were identified on September 12 using a systematic search in EMBASE- and MEDLINE-indexed journals between May 1 and September 11, 2012 (see Appendix A at www.jclinepi.com). Titles and abstracts were screened and then full texts were read to make a final selection. We distinguished between reviews with and without meta-analyses. A meta-analysis was defined as a review in which a pooled estimate for at least one accuracy estimator was reported or, alternatively, in which a summary receiver operating characteristic (SROC) curve was provided.

As this article is about *formal methods* for investigating sources of heterogeneity, as opposed to narrative descriptions of heterogeneity, the primary articles of interest were reviews that contained a meta-analysis. However, as one approach to deal with a high amount heterogeneity is to not pool the results in a meta-analysis, we also performed a brief subsidiary examination of systematic reviews without a meta-analysis to document the reasons review authors provided for not pooling. Reviews on prognostic tests (those used to predict a future condition or event rather than to test for the presence or absence of a current one), testing in animals, individual patient data reviews, conference abstracts, and written in languages other than English were excluded.

2.3. Data extraction from the reviews

The data extraction form was pilot tested by performing double data extraction on a third of the articles (by C.A.N., W.A.v.E., E.E.O., J.A.H.d.G., L.H., and M.M.L.). Discrepancies were discussed, and unclear questions on the form were made more specific. Data extraction was then performed by one researcher (by C.A.N., W.A.v.E., and E.E.O.) using the standardized form and checked by another (by C.A.N., W.A.v.E., or E.E.O.).

Systematic reviews often contain more than one meta-analysis. To prevent the dominance of reviews containing multiple meta-analyses, only information from the main meta-analysis was collected. For objectiveness and clarity in data extraction, the main meta-analysis was defined as the largest group of studies for which a meta-analysis was performed. We thought that the largest meta-analysis was also most likely to have explored sources of heterogeneity. As we were more interested in the range of possibilities for exploring sources of heterogeneity than in precise counts of methods used, we do not think that this selection of the largest meta-analyses will bias our conclusions.

In addition to general review characteristics gathered for the overarching project, information was extracted on the following: whether sources of heterogeneity were explored, the number and type of sources explored, the methods used to explore these sources, how these results were reported, and how conclusions about sources of heterogeneity were made.

For the systematic reviews without a meta-analysis, we extracted the reasons why review authors refrained from calculating pooled estimates. We were particularly interested in seeing if heterogeneity was one of the reasons given for not performing a meta-analysis. When the reviews with a meta-analysis did not explore sources of heterogeneity, information was extracted about the reasons why they did not. What review authors reported about how they intended to make a decision on whether to explore sources of heterogeneity was also recorded.

When counting the number of potential sources of heterogeneity explored (which we refer to hereafter as ''factors''), the types of factors were counted, rather than the number of subgroups or strata of those factors. For example, if threshold effects were explored and summary estimates were presented for several cutoff points, threshold was only counted as one factor. The relationship (ratio) between the number of factors explored and the number of primary studies in the review was analyzed, as well as the relationship between the number of factors explored and the method used to perform the exploration.

The factors explored were categorized as clinical, quality (ie, study design characteristics), index test related, or ''other.'' Explanation of some of the quality-related factors explored can be found in QUADAS-2, a revised tool for the quality assessment of diagnostic accuracy studies [9]. Publication year of the target disease under study was categorized under the category ''other,'' because it is often difficult to know what it truly measures. For example, over time, technological advances may result in the accuracy of an imaging test improving, but at the same time, the patient spectrum could also change. In such a situation, publication year could be categorized as a clinical or an index test–related source of heterogeneity. In addition to categorizing the factors, it was also noted whether continuous factors were categorized or sum scores (scores which summarize information about several factors into a single value) were used.

The methods used to explore sources of heterogeneity were classified into three categories: sensitivity analysis, stratification, and metaregression. We defined sensitivity analysis as the exclusion of one or more studies from the meta-analysis for the purpose of seeing how the summary estimates in the reduced group differ from the overall estimate [10]. Stratified analysis was defined as the calculation of summary estimates for subgroups defined by a particular factor (eg, providing separate estimates of sensitivity and specificity for each gender). Metaregression was defined as the entering of a factor or factors into a metaregression model as coefficients to explore how they influenced the summary estimates.

Because there are different ways to come to conclusions about whether a specific factor is a source of heterogeneity, what the authors reported about how they made this conclusion was recorded. Additionally, it was noted whether the sources were tested statistically, whether the results were presented per subgroup, and whether the reduction in heterogeneity or remaining heterogeneity (within a subgroup compared with all groups combined) was reported.

Information was extracted on what authors reported in the abstract and conclusion about the exploration of sources of heterogeneity. Studies that discussed this in either the abstract or the conclusion were considered to place a high importance on this topic, as these are the sections in which the most important findings are typically discussed and which most readers often base their own conclusions on [11].

As methodological reviews should go beyond only describing what has been done to provide assistance to researchers [12], a list of items for researchers to consider and report on when exploring sources of heterogeneity in a systematic review of diagnostic studies was developed. The domains in the list parallel data extraction (ie, whether to explore sources of heterogeneity, selection of factors to explore, methods of exploration, and presentation and interpretation of results) and the contents were inspired by the variety of approaches observed in the reviews.

3. Results

3.1. Search results

After exclusion of duplicates, the search resulted in a total of 1,273 hits. On screening of titles and the exclusion of

articles that were only conference abstracts, 1,058 articles were excluded. The full text of the remaining 89 potentially relevant articles was reviewed to determine whether they met the inclusion criteria. After this process, 65 systematic reviews were identified of which 53 contained at least one meta-analysis. Appendix B (see at www.jclinepi.com) contains the search results details, and Appendix C (see at www.jclinepi.com) contains a list of the included reviews.

3.2. General characteristics of the reviews

Of the 12 systematic reviews that did not perform a meta-analysis, eight stated that they did not do so because there was too much heterogeneity. Other reasons given for not performing a meta-analysis were low quality of the primary studies ($n = 4$), too few primary studies ($n = 2$), and studies having different cutoffs ($n = 1$).

The general characteristics of the 53 reviews that contained a meta-analysis can be found in Table 1. The meta-analyses contained a median of 14 primary studies (interquartile range [IQR], 9.5–18.5). Most of the reviews were on imaging tests (60%), a large percentage was on laboratory tests (26%), and a few were on clinical examination procedures (14%). Over half of the meta-analyses investigated more than one index test, and most of these contained a comparative question.

More than half of the reviews that contained a meta-analysis tested for heterogeneity using Cochran Q test ($n = 28$), and more than half of the reviews measured it using I^2 ($n = 31$) [13,14]. Very few presented the between-study variance estimate (τ^2) from a random-effects model ($n = 7$), and even fewer interpreted this for the reader by presenting prediction intervals or ellipses ($n = 3$). Prediction regions show the range of values where the true value from new comparable study is likely to be found [2]. When obtaining summary estimates of test accuracy, only about a third used a more advanced hierarchical bivariate model ($n = 13$) [15,16], whereas approximately half used an SROC according to Moses and Littenberg ($n = 24$) [17], and the rest only undertook univariate pooling ($n = 13$).

Almost all the reviews with a meta-analysis performed a quality assessment of the primary studies ($n = 49$). The vast majority of studies used the formal tool QUADAS (a quality assessment tool for diagnostic accuracy studies; $n = 40$) [18]. As the newer version of this tool, QUADAS-2, was only introduced at the end of 2011, it is logical that it was only used in one of the reviews [19].

3.3. Number and type of sources of heterogeneity explored

Forty of the fifty-three reviews containing a meta-analysis formally explored sources of heterogeneity. There was a large spread in the number of factors that were explored as potential sources of heterogeneity (Fig. 1).

Table 1. Characteristics of all reviews containing a meta-analysis ($n = 53$) and of the subset in which heterogeneity was investigated statistically ($n = 40$)

Number of primary studies, median (IQR)	14 (9.5–18.5)
Size of primary studies, median (IQR)	87 (45–182)
Type of study, *n* (%)	
Image	32 (60)
Laboratory	15 (28)
Clinical examination	6 (11)
Meta-analyses looking at more than one index test, *n* (%)	31 (58)
Two tests	14 (26)
Three to six tests	16 (30)
More than six tests	1 (2)
Contained a comparative question, *n* (%)	21 (40)
Testing and measuring heterogeneity, *n* (%)	
Cochran Q test	28 (53)
I^2	31 (58)
τ^2	7 (13)
Prediction intervals, ellipses, or bands	3 (6)
Method(s) for conducting the meta-analysis, *n* (%)	
Univariate analysis only	13 (26)
SROC (Moses–Littenberg): linear regression D on S	24 (48)
HSROC (Rutter and Gatsonis): accuracy, scale, and threshold parameter	5 (5)
Bivariate random-effects model (Reitsma): random effects sensitivity and specificity	13 (26)
Studies performing a quality assessment, *n* (%)	
QUADAS	40 (75)
QUADAS-2	1 (2)
STARD	5 (9)
Other or own instrument	5 (9)
No quality assessment	4 (8)
Studies investigating sources of heterogeneity statistically, *n* (%)	40 (75)
Methods for investigating sources of heterogeneity ($n = 40$), *n* (%)[a]	
Sensitivity analysis	12 (30)
Stratified analysis	25 (63)
Metaregression	19 (48)
Number of potential sources of heterogeneity explored/number of primary studies in meta-analysis ($n = 40$), median (IQR)	0.21 (0.09–0.46)
Number of reviews identifying sources of heterogeneity ($n = 40$), *n* (%)	
At least one	29 (73)
More than one	8 (20)

Abbreviations: IQR, interquartile range; QUADAS, a quality assessment tool for diagnostic accuracy studies; HSROC, Hierarchical summary receiver-operator curves; STARD, STAndards for the Reporting of Diagnostic accuracy studies.

[a] These numbers do not add up to 40 because some studies used more than one of the methods.

The median ratio of factors explored relative to the number of primary studies contained in a meta-analysis was approximately one factor for every five primary studies. In 33 of the 40 meta-analyses exploring sources of heterogeneity (80%), more than one factor was explored for every 10 studies included. Note that we only counted factors that were actually formally explored. Authors often stated in the methods that they would look at many factors as sources of heterogeneity, but for various reasons (eg, too few studies

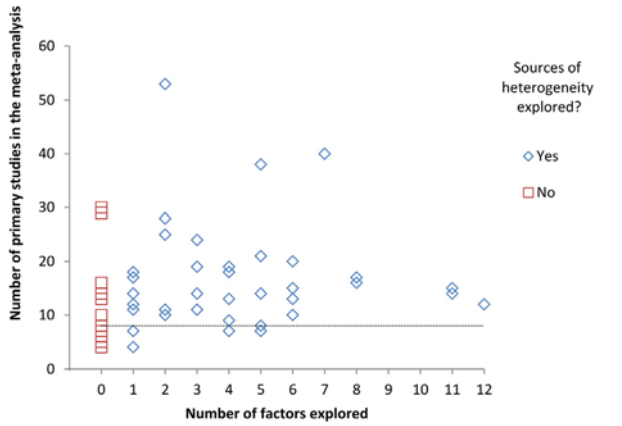

Fig. 1. Relationship between the number of potential sources of heterogeneity explored (factors) and the number of primary studies in the meta-analysis. The dashed line is drawn at eight studies, the median number of studies in meta-analyses that did not explore sources of heterogeneity. One outlying study with 114 primary studies, which explored 11 factors of heterogeneity, was excluded to improve readability.

in the meta-analysis or poor reporting on the factor of interest), some of them could not be explored.

A breakdown of the number of meta-analyses investigating particular types of factors can be found in Table 2. There did not appear to be a difference in the frequency

Table 2. Sources of heterogeneity explored (n = 40)

Category	Factor	Number of studies investigating factor (n = 40)
Clinical factors		25
	Age	7
	Sex	2
	Spectrum or other clinical-related factors	18
	Prevalence	8
Study quality items[a]		25
	Blinding	7
	Sample size	9
	Reference test	9
	Verification biases	5
	Quality score	5
	Prospective vs. retrospective design	8
	Consecutive vs. nonconsecutive enrollment	4
	Other quality items	8
Test or threshold		20
	Variation of the index test	15
	Index test threshold	7
Other		
	Publication year	9
	Other not classified[b]	4

[a] Explanation of how some of these quality items can introduce bias can be found in QUADAS-2, a revised tool for the quality assessment of diagnostic accuracy studies [9].

[b] Other: studies that appeared to be outliers based on visual assessment of the receiver operating characteristic curve (n = 1) percentage of index test positives (n = 1) and a leave-one-out sensitivity analysis (n = 2).

in which any of the categories of factors (ie, clinical, quality (study design characteristics), or index test related) were explored. The factors categorized under the "other" category of sources of heterogeneity were the percentage of index test positives (n = 1), studies that appeared to be outliers based on visual assessment of the ROC curve (n = 1), and a leave-one-out (outlier) analyses (n = 2).

While some studies reported that they prespecified what factors they would explore, some reviews may have decided which factors to explore based (partially) on visual exploration of the results presented in forest or ROC plots. This was not information that we extracted from the meta-analyses, as it is often impossible to tell whether authors selected factors before or after the gathering of the primary study results. However, we did observe a difference in approaches from comments made by the authors. For example, one reviewer reported that "heterogeneity was evaluated visually through observed differences between study characteristics and methodologies, and by examining for substantial difference in the sensitivities and specificities on the forest plot. Studies that demonstrated considerable heterogeneity were excluded from the meta-analysis." [20].

Of the 25 reviews in which continuous factors were explored (such as mean age, publication year, or prevalence), 15 studies dichotomized these factors. In the other 10, it was unclear how these factors were explored. Only 3 of the 15 reviews that had dichotomized the factor of interest provided the reason behind the chosen cutoff. Of the 25 reviews that explored quality items, quality sum scores were explored as a potential factor of heterogeneity in five reviews.

3.4. Methods of investigating sources of heterogeneity

Of the 40 studies investigating heterogeneity, 12 performed sensitivity analysis, 25 stratification, and 19 metaregression (Table 3). Fifteen studies used two of these methods and one used all the three.

The number of factors explored varied by methods for investigating heterogeneity. Sensitivity analysis was only conducted on one or two factors per meta-analysis, whereas a median of three factors were explored using stratification, and four through metaregression. Comparing stratification to metaregression, a high numbers of factors (ie, more than six) were only explored using metaregression. Although the number of factors explored varied between the methods, there was no difference observed by method type in terms of the number of primary studies or the total number of subjects in the meta-analysis.

In the studies using metaregression, it was often unclear whether they had explored factors one by one (eg, by fitting multiple models) or whether they entered multiple factors into a single model. In 7 of the 18 studies that had looked at more than one potential factor, it was reported that multiple factors were put into the model at the same time.

Table 3. Comparison of methods of investigating sources of heterogeneity (n = 40)

Meta-analysis characteristics	Only a narrative exploration of heterogeneity (n = 13)	Explored sources of heterogeneity statistically (n = 40)	Sensitivity (N = 12)[a]	Stratification (N = 25)[a]	Metaregression (N = 19)[a]
Number of factors explored, median (interquartile range [IQR])	—	—	1 (1–1)	3 (1–4)	4 (3–6)
1	—	—	10	10	1
2–3	—	—	2	6	6
4–5	—	—	0	9	5
≥6	—	—	0	0	7
Study size, median (IQR)					
Number of primary studies included	8 (5–15)	14.5 (11–19)	14.5 (12.25–22.25)	15 (11–18)	15 (11–20)
Number of subjects in meta-analysis	560 (174–1,716)	2,106 (636–6,495)	2,150.5 (766–8,635)	1,725 (493–4,828)	2,576 (1,112–13,662)
Ratio of number of factors explored to the number of studies in the meta-analysis, median (IQR)	—	0.21 (0.09–0.46)	0.07 (0.06–0.12)	0.15 (0.07–0.26)	0.27 (0.15–0.50)
Authors concluded that there was significant or meaningful true heterogeneity	8 (62%)	27 (68%)	—	—	—

[a] The numbers do not add up to 40 because some studies used more than one method: three studies performed sensitivity analysis and stratification, four sensitivity analysis and metaregression, eight metaregression and stratification, and one study used all three methods.

However, in the other 11 studies, it was unclear how factors had been entered and removed from the model.

3.5. Reviews that did not examine sources of heterogeneity

Although most meta-analyses attempted to explain the variety in study results in a descriptive manner, one-fourth (n = 13) did not explore sources of heterogeneity formally. Meta-analyses that did not explore sources of heterogeneity were not very different from those that did. Overall, they were slightly smaller in terms of the number of primary studies and participants included in the meta-analysis (Table 3). Still, several studies that explored sources of heterogeneity were smaller than those that did not (Fig. 1). Authors concluded that there was significant heterogeneity in about two-thirds of the meta-analyses in both groups (those exploring and those not exploring sources of heterogeneity; Table 3).

Of the 13 meta-analyses that did not report on the formal exploration of sources of heterogeneity, only one author reported the reason why it did not, namely, that there were too few studies. In the methods section, four articles (of the 53 meta-analyses) announced that the results of the tests for true heterogeneity would influence the decision of whether to explore sources of heterogeneity.

3.6. Interpretation and presentation of results

Although many reviews explored sources of heterogeneity (n = 40), the methods to which they came to conclusions about sources of heterogeneity, the thoroughness to which they reported their results, and the importance that was given to the findings of this exploration varied (Table 4).

In total, only 11 (28%) gave a clear description of how they defined sources of heterogeneity. A variety of methods for defining a significant source of heterogeneity was observed, such as comparing the confidence intervals of

Table 4. Meta-analyses differ in how they analyze, report, and present their conclusions about the sources of heterogeneity (n = 40)

Analysis and reporting			
Defined significant source of heterogeneity	11		
Presented results per subgroup	34		
Reduction in heterogeneity was reported	6		
Conclusions	**In abstract**	**In conclusions**	**In either**
No mention of sources of heterogeneity	16	5	3
Unable to *explore* what causes the heterogeneity	0	6	6
Unable to *conclude* what causes the heterogeneity	4	8	10
Identified factors as sources of heterogeneity	8	24	25
Presented subgroup results	14	Not applicable[a]	14
Interpreted subgroup results	13	Not applicable[a]	13

[a] This information was not applicable as most studies present the results in the results section, not the conclusion.

subgroups, looking at the *P*-value of the regression coefficient in metaregression, and testing if "true" heterogeneity was (still) present within the subgroups. Twenty-nine studies (73%) identified at least one source of heterogeneity and eight (20%) identified more than one source. Of these 29 studies, only eight (28%) explained how they came to this conclusion.

Some researchers only performed the exploration of sources of heterogeneity without presenting the results in a form that was easy for readers to interpret (ie, by only performing statistical testing or presenting coefficients from metaregression; $n = 6$, 15%), whereas others (also) presented stratified results for at least one factor explored ($n = 34$, 85%).

Only 3 (8%) of the 40 meta-analyses that had explored sources of heterogeneity did not mention anything about this exploration, or the findings thereof, in either the abstract or the conclusion (the main findings). Of the 29 studies identifying a source of heterogeneity, 25 (86%) reported this finding in the main findings. On the other hand, 14 (35%) authors reported in the main findings that they were either unable to explore (particular) sources of heterogeneity or unable to come to a conclusion about the cause of heterogeneity. When reporting the findings in the abstract, authors usually ($n = 13$, 93%) presented and gave a clinical or methodological explanation for the subgroup results.

4. Discussion

4.1. Strengths and limitations of this review

In addition to documenting how sources of heterogeneity are currently being explored, this review goes one step further than existing reviews (in the following discussion) to provide a list of items that researchers can consider and report on when investigating sources of heterogeneity [8,21]. Although we do not provide formal guidance, the list of items we provide will be helpful to researchers in that it raises awareness of the various options available. This list can be found in Table 5.

Although our sample size of meta-analyses was somewhat smaller than prior methodological studies on systematic reviews of diagnostic studies [8,21] we think that it was sufficiently large. We think that our sample size was large enough because our goal was to get an idea of the range of approaches used rather than to precisely estimate how many studies took each approach to investigating sources of heterogeneity [11].

4.2. Whether to explore sources of heterogeneity statistically

The first decision to make when looking into why results vary between primary studies is whether to explore the

Table 5. Summary items to consider and report on when exploring sources of heterogeneity

Domain	Key items to consider and report on
Whether to explore sources of heterogeneity	• Consider and report how this decision will be made • Report why the exploration was not possible
Selecting potential sources of heterogeneity to explore	• Consider whether to limit the number of factors explored • Consider and report on how potential sources will be selected: ○ Motivated analysis (a few factors thought to be of most particular clinical interest or cause severe bias) or exploratory analysis (many available factors) ○ A priori or a posteriori selection of factors • Consider exploring individual quality items instead of quality sum scores • Consider whether each factor is a patient- or study-level characteristic. ○ When it is a patient-level factor, consider whether subgroup estimates can be extracted (eg, separate estimates for male and female) as opposed to study-level characteristics (eg, percentage male and female)
Methods of exploring sources of heterogeneity	• Consider, for each factor being explored, what method to use to explore sources of heterogeneity: ○ Sensitivity, stratified analysis, or (bivariate) metaregression • If there are two main outcomes of interest in the study (ie, sensitivity and specificity), consider using bivariate metaregression • When performing (bivariate) metaregression: ○ Consider and report the form of the factors being explored (categorical or continuous). If factors are categorized, report the cutoff points and reasoning behind them. ○ Consider and report how factors are entered into the model (a separate model for each factor, all factors in the same model, and so forth)
Interpretation and Presentation of results	• Consider and report how conclusions are drawn about what is a significant source of heterogeneity • Consider whether reporting stratified results will help interpretation • Before concluding that a particular factor causes heterogeneity, consider what other closely related factors could also have caused it • Consider and report why factors identified as sources of heterogeneity could cause heterogeneity

potential factors causing this variability in a formal statistical manner as opposed to simply providing a narrative description. It is important to acknowledge that there are several insurmountable barriers to formally investigating sources of heterogeneity. In addition to a small number of primary studies included in the review, poor reporting in the primary studies or high similarity of the studies in terms of study design and study population can make investigating sources of heterogeneity difficult [7].

Although it makes sense that there is no need to explore the source of heterogeneity if there is no true heterogeneity detected, defining true heterogeneity is challenging. Authors may judge whether there is heterogeneity from visual inspection of the forest or ROC plots. Although viewing data can provide insights into the variability between studies, it is subjective and formal inferences about the presence of true heterogeneity can only be made based on statistical tests and measurements [3,22].

That said, statistical tests and measurements of heterogeneity also have their pitfalls. Tests for heterogeneity, such as the Cochrane Q statistic have low power for the typical review of diagnostic tests in which often few and also relatively small studies are included [1,23]. Likewise, the confidence interval around I^2 will be very large when there are few studies, meaning that there is large uncertainty about heterogeneity. This high degree of uncertainty makes the I^2 difficult to use when making a decision about whether to explore sources of heterogeneity, regardless of the chosen cut point [22]. Furthermore, the I^2 does not take into account the correlation between sensitivity and specificity.

The bivariate random-effects model provides metrics for heterogeneity that take into account the correlation between sensitivity and specificity [15]. This model provides three parameters: between-study variance (τ^2) in sensitivities, in specificities, and the covariance between them. The combination of these three metrics makes it possible to examine total study variance in sensitivity or specificity as well as conditional variance (the variance in sensitivities at a fixed value of specificity or vice versa). However, the τ^2 values are difficult for authors to interpret. More guidance is needed on interpreting the tests and measurements for true heterogeneity in diagnostic studies.

4.3. Selecting potential sources of heterogeneity to explore

Overall, the included reviews explored a high number of potential sources of heterogeneity compared with the number of studies in the meta-analysis (median, 1:5). We caution against the use of testing a high number of factors compared with the number of studies in the meta-analysis to avoid the well-known problem of multiple testing. When too many factors are tested, the risk of incorrectly concluding a factor causes (some of) the heterogeneity increases. There is no recommended ratio of the number of factors to number of studies which can be explored in a meta-analysis; however, a common rule of thumb in regression analysis is that for every covariate (in this case factor), there should be at least 10 observations (in this case primary studies) [24,25].

It is difficult to translate this rule to bivariate metaregression, as the initial model itself, without any covariates, already has five parameters that need to be estimated and each covariate that is explored adds two additional parameters to be estimated, instead of only one as is the case in regression analysis.

Although it is difficult to judge exactly how authors choose which sources to explore, two general approaches to selecting sources of heterogeneity to explore were identified: motivated and exploratory. The motivated approach is to carefully select a few factors to explore for which one has reasons to believe that they may lead to differences in accuracy. The exploratory approach is to explore many potential factors regardless of whether there is a strong reason to believe that each factor could influence test accuracy. It is helpful to communicate to the reader whether the choice of factors was motivated or exploratory as well as whether the factors were selected before or after observing the results [26,27].

Some of the factors explored are categorical by nature, but many are continuous, such as age, prevalence of disease, or publication year. Careful thought should be given to whether to categorize factors, and it is important to mention the cutoff value(s) as well as the reasoning behind them [28]. Additionally, when performing a metaregression, it is important to consider and report whether factors are explored one by one or multiple factors were entered into the model at once.

Sometimes authors calculate a quality sum score to get a better feel for which studies are of higher quality than others. Because sum scores give equal weighting to unequal factors, exploring sum scores as factors of heterogeneity is generally discouraged [29]. Post hoc exclusion of studies based on only visual inspection of the ROC plot analysis or a leave-one-out analysis is discouraged as well.

Sources of heterogeneity can be divided into two distinct groups: those that relate to characteristics of the patients included in a diagnostic study (eg, gender, age, or severity of symptoms) and those that characterize the primary studies (eg, whether all patients received the same reference standard or the year of publication). Exploring patient-level characteristics in a review brings additional challenges. In general, the power for examining patent-level characteristics is low unless the individual studies report separate two-by-two tables for the different categories of that factor. When such results are not available, the only approach left is to use study-level summary measures, such as mean age or percentage of males.

The use of study-level summary measures to investigate patient-level characteristics is problematic. For example, if there was a true difference in diagnostic accuracy between genders, this source of heterogeneity would

go undetected if each study contained an equal number of males and females. Individual patient data meta-analyses are much more equipped for examining differences in accuracy that relate to patient-level characteristics [30]. In general, review authors should be cautious when examining the relevance of patient-level characteristics in their review, unless primary studies report stratified data for that factor.

4.4. Methods of exploring sources of heterogeneity

Because it is not necessary to choose the same method of exploration for each potential source of heterogeneity, authors may consider the individual factors they are investigating before choosing an appropriate method. If the main interest is in a specific group of studies or one wants to study the robustness (ie, through a leave-one-out analysis) of the meta-analyses results, a sensitivity analysis can be considered. Visual inspection of the ROC plot can be used to detect outliers or overly influential studies, but it is subjective, especially when study size is not represented. A good reason to do a sensitivity analysis is to get an estimate from only the high-quality studies by excluding the low-quality studies or studies with poor reporting. After all, the high-quality studies are the ones on which clinical inferences can best be drawn.

If the interest is in comparing summary estimates between groups (eg, seeing if the test performs differently in primary vs. secondary care), stratified analysis or metaregression is a logical choice. Stratified analysis is more focused on comparing the estimates between the subgroups, whereas metaregression is focused on whether the factor is associated with a difference in accuracy between the groups. However, results from metaregression for categorical factors can also be presented in a stratified manner.

In metaregression, it is possible to explore multiple factors simultaneously and to explore factors in their continuous form. Authors should report details on how factors were entered into the model and in what form. As authors performing metaregression explored more factors than those doing stratified analysis or sensitivity analysis, it seems like metaregression is being used more often for exploratory rather than confirmative analysis. However, the problem of multiple testing cannot be avoided using multiple regression as opposed to sensitivity analysis or stratified analysis.

4.5. Interpretation and presentation of results

Regardless of the chosen method for investigating sources of heterogeneity, it is important that authors define and report how they will conclude whether a factor is a significant source of heterogeneity. Researchers may consider presenting stratified results when a factor is detected to help convey the clinical or methodological relevance.

Although much work can go into investigating sources of heterogeneity, the effect of a factor may not always be identified. On the other hand, when a source of heterogeneity is identified statistically, caution should be exercised against jumping to the conclusion that the factor is actually causing the heterogeneity. If factors are closely related to each other, it is often impossible to determine which factor is causing the heterogeneity. Instead of just attributing heterogeneity to a factor, one should try to explain why that factor could be causing heterogeneity. When an identified sources of heterogeneity is a quality item, it is particularly important to explain what that item is as readers may not be familiar with them [12]. Ultimately, the exploration of heterogeneity is performed with the hope that factors will be identified that are relevant for current clinical practice or future research.

5. Conclusion

In this review, we found that methods for exploring sources of heterogeneity in meta-analyses of diagnostic studies vary widely between meta-analyses. Based on the variety in methods observed, we developed a list of items to consider and report on. While waiting for formal guidance to be developed, this list can be used by researchers in the meantime to improve the way that they explore sources of heterogeneity and report findings of this exploration in meta-analyses of diagnostic studies.

Appendix

Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.jclinepi.2014.05.018.

References

[1] Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med 2002;21:1539–58.
[2] Riley RD, Higgins JP, Deeks JJ. Interpretation of random effects meta-analyses. BMJ 2011;342:d549.
[3] Deeks J, Bossuyt P, Gatsonis C. Cochrane handbook for systematic reviews of diagnostic test accuracy version 1.0. The Cochrane Collaboration 2010. Available at http://srdta.cochrane.org. Accessed August 1, 2013.
[4] Cochrane handbook for systematic reviews of interventions. Chichester, UK: Wiley-Blackwell; 2008.
[5] Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 7. Rating the quality of evidence—inconsistency. J Clin Epidemiol 2011;64:1294–302.
[6] Schunemann 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;336:1106–10.
[7] Higgins JP, Green S. The Cochrane handbook for systematic reviews of interventions. Version 5.1.0 ed. 2011.
[8] Dahabreh IJ, Chung M, Kitsios GD, Terasawa T, Raman G, Tatsioni A, et al. Comprehensive overview of methods and reporting

of meta-analyses of test accuracy. Rockville, MD: Agency for Healthcare Research and Quality (US); 2012.
[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;155:529–36.
[10] Porta M. A dictionary of epidemiology. 5th ed. New York: Oxford University Press; 2008.
[11] Lilford RJ, Richardson A, Stevens A, Fitzpatrick R, Edwards S, Rock F, et al. Issues in methodological research: perspectives from researchers and commissioners. Health Technol Assess 2001;5:1–57.
[12] Zhelev Z, Garside R, Hyde C. A qualitative study into the difficulties experienced by healthcare decision makers when reading a Cochrane diagnostic test accuracy review. Syst Rev 2013;2:32.
[13] Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003;327:557–60.
[14] Rucker G, Schwarzer G, Carpenter JR, Schumacher M. Undue reliance on I(2) in assessing heterogeneity may mislead. BMC Med Res Methodol 2008;8:79–88.
[15] Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. J Clin Epidemiol 2005;58:982–90.
[16] Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Stat Med 2001;20:2865–84.
[17] Moses LE, Shapiro D, Littenberg B. Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. Stat Med 1993;12:1293–316.
[18] Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC Med Res Methodol 2003;3:25.
[19] Schueler S, Schuetz GM, Dewey M. The revised QUADAS-2 tool. Ann Intern Med 2012;156:323–4.
[20] Smith TO, Drew BT, Toms AP. A meta-analysis of the diagnostic test accuracy of MRA and MRI for the detection of glenoid labral injury. [Review]. Arch Orthop Trauma Surg 2012;132(7):905–19.
[21] Dinnes J, Deeks J, Kirby J, Roderick P. A methodological review of how heterogeneity has been examined in systematic reviews of diagnostic test accuracy. Health Technol Assess 2005;9:1–113. iii.
[22] Ioannidis JP, Patsopoulos NA, Evangelou E. Uncertainty in heterogeneity estimates in meta-analyses. BMJ 2007;335:914–6.
[23] Ioannidis JP. Interpretation of tests of heterogeneity and bias in meta-analysis. J Eval Clin Pract 2008;14:951–7.
[24] Peduzzi P, Concato J, Feinstein AR, Holford TR. Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. J Clin Epidemiol 1995;48:1503–10.
[25] Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 1996;49:1373–9.
[26] Thompson SG, Higgins JP. How should meta-regression analyses be undertaken and interpreted? Stat Med 2002;21:1559–73.
[27] Higgins JP, Thompson SG. Controlling the risk of spurious findings from meta-regression. Stat Med 2004;23:1663–82.
[28] Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. Stat Med 2006;25:127–41.
[29] Whiting P, Harbord R, Kleijnen J. No role for quality scores in systematic reviews of diagnostic accuracy studies. BMC Med Res Methodol 2005;5:19.
[30] ter RG, Bachmann LM, Kessels AG, Khan KS. Individual patient data meta-analysis of diagnostic studies: opportunities and challenges. Evid Based Med 2013;18(5):165–9.