PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Diffusion-weighted magnetic resonance imaging in differentiating malignant from benign thyroid nodules: A meta-analysis
AUTHORS	Chen, Lihua; Xu, Jian; Bao, Jing; Huang, Xuequan; Hu, Xiaofei; Xia, Yunbao; Wang, Jian

VERSION 1 - REVIEW

REVIEWER	Jin Wei QIANG Department of Radiology, Jinshan Hospital, Fudan University, Shanghai, China
REVIEW RETURNED	11-Jun-2015

GENERAL COMMENTS	This paper reports on the value of DWI for differentiation of malignant nodules from benign thyroid nodules by meta-analysis. The topic is valuable for clinical practice, but there are several issues that must be clarified to better understand the methods and interpret the results. 1. Inclusion criteria is not fully defined. According to the existing criteria, low quality research reports and selecting bias cannot be reliably excluded. 2. There is no mention of histopathological types of thyroid nodules which are in great variation according to subtypes of inflammatory nodules, adenoma, carcinoma, lymphoma, etc. 3. For primary research quality assessment, interobserver variability should be performed. 4. There is a wide variation of b values from 100 to 1000. I suggest to reevaluate the value of DWI according to the range of b value (low and high b value group). 5. Lack of description of ADC reproducibility was the critical point that diminish the value of results in this study. 6. Evidence-based decision cannot be made according to the article due to the current inaccurate, biased methodology. 7. There were innumerable grammatical and spelling errors in the paper.

REVIEWER	Mitsuhiko Nakahira Saitama Medical University, International Medical Center, Saitama, Japan
REVIEW RETURNED	18-Jun-2015

GENERAL COMMENTS	In this study, the authors have done systematically review of the studies related to DWI in differentiating between malignant and benign thyroid nodules.

Comments:
1. I identified six authors listed in this article. In Methods, the authors stated that database search was done by two independent observers. Who were two observers responsible for this work? Please state it more specifically in both stages of literature search and selection of article. In addition, in the next stage, quality assessment and data extraction, the authors stated that the aforementioned three authors extracted data from the selected reports. However, I can identify only two observers who worked in the previous stages so far. Who is the third author?
2. In Meta-analysis, the authors used several statistical techniques to assess heterogeneity of quality of articles. Because I do not have sufficient statistical knowledge and experience to justify whether such techniques are appropriate for use, I marked a item that this paper requires specialist statistical review.
3. The number of articles in a square frame of Fig. 1 is not in accordance with description in the text.
4. This study that includes the number of papers more than twice, compared to the precedent meta-analysis published in 2014 shows the same result that DWI have a potential to differentiate malignant from benign thyroid nodules as ever. Additionally, although subgroup analyses based on adopted b value, magnetic field, study design and blinded study have been completed to identify impact factors on the accuracy of DWI, there were no significant impact factors correlated with accuracy. In summary, to use DWI in discriminating between malignant thyroid nodules and benign nodules, this report gives some more favorable information than ever before.

VERSION 1 – AUTHOR RESPONSE

Response to Reviewer 1:
1. Inclusion criteria is not fully defined. According to the existing criteria, low quality research reports and selecting bias cannot be reliably excluded.
To perform a full review, the inclusion criteria of the studies in our meta-analysis is relaxed. In that way, there are some factors that can result in accuracy estimates amongst different test accuracy studies in a review, such as design (prospective or retrospective), the lesions size, reference standard, enrolment, magnetic field strengths, patient spectrum, sequences. Thus, meta-regression and subgroup analyses were performed to detect the factors that impact heterogeneity.
We did not exclud low quality research reports, because the quality of studies can result in accuracy estimates amongst different test accuracy studies. As the quality assessment of these 15 studies was moderate in according to QUADAS-2 items, it hard to do subgroup analysis on quality of studies (low quality or high quality). The results of the distribution of the study design are shown in Fig 2.

2. There is no mention of histopathological types of thyroid nodules which are in great variation according to subtypes of inflammatory nodules, adenoma, carcinoma, lymphoma, etc.
We are quite agreeing with you. We intended to evaluate the histopathological types of thyroid nodules, however, all the 15 articles lacked sufficient data to complete a 2×2 contingency table.

3. For primary research quality assessment, interobserver variability should be performed.
We are quite agreeing with you.

4. There is a wide variation of b values from 100 to 1000. I suggest to reevaluate the value of DWI according to the range of b value (low and high b value group).
We intended to evaluate the b value to low b value group and high b value group. However, more studies reported data of 500 s/m2 than 300 s/m2 or 1000 s/m2, it hard to decide whether put 500

s/m2 to low b value group or high b value group.

5. Lack of description of ADC reproducibility was the critical point that diminish the value of results in this study.
We are quite agreeing with you. Most studies included in our meta-analysis lacked the description of ADC reproducibility. We also added this limitation in our manuscript.

6. Evidence-based decision cannot be made according to the article due to the current inaccurate, biased methodology.
The methodology of our meta-analysis was according to PRISMA Statement. Please see the PRISMA 2009 checklist in Checklist S1.
As some inherent limitations exist in our study, we had tone down our evaluation of the main study results in Conclusions section.
7. There were innumerable grammatical and spelling errors in the paper.
This manuscript had been sent to American Journal Experts and edited by a native speaker of English. The AJE certificate of our manuscript is shown in Fig 1.

Response to Reviewer 2:
1. I identified six authors listed in this article. In Methods, the authors stated that database search was done by two independent observers. Who were two observers responsible for this work? Please state it more specifically in both stages of literature search and selection of article. In addition, in the next stage, quality assessment and data extraction, the authors stated that the aforementioned three authors extracted data from the selected reports. However, I can identify only two observers who worked in the previous stages so far. Who is the third author?
In our selection of articles section, two authors (Lihua Chen and Jian Xu) initially screened the titles and abstracts of the search results and retrieved the full text of all potentially relevant reports. Next, the authors independently reviewed all relevant reports, according to the predefined inclusion criteria. Disagreements were resolved by a consensus or third author (Jian Wang) arbitration who assessed all of the involved items. The majority opinion was used to determine whether a particular study met the selection criteria.
In the quality assessment and data extraction, the aforementioned three authors (Lihua Chen, Jian Xu, and Jian Wang) extracted data from the selected reports.

2. In Meta-analysis, the authors used several statistical techniques to assess heterogeneity of quality of articles. Because I do not have sufficient statistical knowledge and experience to justify whether such techniques are appropriate for use, I marked a item that this paper requires specialist statistical review.
We had used the same statistical techniques and published some papers, for example:
1. Chen L, et al. Direct comparison of cardiovascular magnetic resonance and single-photon emission computed tomography for detection of coronary artery disease: a meta-analysis. PLoS One. 2014, 9(2):e88402.
2. Lihua Chen, et al. Magnetic resonance imaging with gadoxetic acid disodium for the detection of hepatocellular carcinoma: a meta-analysis of 18 studies. Acad Radiol. 2014, 21(12):1603–1613.

3. The number of articles in a square frame of Fig. 1 is not in accordance with description in the text.
We are sorry to make a mistake and had revised it.

4. This study that includes the number of papers more than twice, compared to the precedent meta-analysis published in 2014 shows the same result that DWI have a potential to differentiate malignant from benign thyroid nodules as ever. Additionally, although subgroup analyses based on adopted b value, magnetic field, study design and blinded study have been completed to identify impact factors

on the accuracy of DWI, there were no significant impact factors correlated with accuracy. In summary, to use DWI in discriminating between malignant thyroid nodules and benign nodules, this report gives some more favorable information than ever before.

In our study, eight additional references were included that were not present in the aforementioned meta-analysis, which make our research more persuasive.

As some inherent limitations exist in our meta-analysis and the aforementioned meta-analysis: (a) These studies are from Asian countries, while no studies were from Europe or North; (b) Most studies lacked the description of ADC reproducibility; (c) Quality assessment of these studies was moderate. Based on these limitations, we had tone down our evaluation of the main study results, which make our evaluation more objective and accuracy.

Compared with previous meta-analysis, we presented a new point that using a high b value may provide better results. It provide an additional value to possible methodological improvements.

Response to Reviewer 3:
Reviewer: 3
Reviewer Name Steven Lane
Institution and Country Department of biostatistics
University of Liverpool
England
Please state any competing interests or state 'None declared': None declared

Please leave your comments for the authors below
Methods
Page 7: Line 9: for to systematically review - grammatical error
We are sorry to make a mistake and had revised it.

Line 17: Estimates might be better than results
We had revised it.

Results
Page 9: Line 18: P values are only for test the regression coefficient different from zero, does not provide evidence that correlation is significant.
Threshold effect could be one important extra source of variation in meta-analysis of diagnostic accuracy: the studies included may have used, explicitly or implicitly, different thresholds to define positive and negative test results. One can test for this threshold effect in two ways: i) plotting sensitivity and specificity on an ROC plane. If such a threshold effect exists, the points will show a curvilinear pattern; ii) calculating the Spearman correlation coefficient between sensitivity and specificity. If the threshold effect exists an inverse correlation appears [1].
We used Meta-Disc version 1.4 to assess the threshold effect. With the threshold option, Meta-DiSc computes the Spearman correlation coefficient between sensitivity and specificity and its p-value. With plot option, it can plot sensitivity and specificity on an ROC plane.
If there is a threshold effect, the best summary of study results is an ROC curve rather than a single pair of pooled indexes, there are two methods of fitting the ROC curve: symmetrical and asymmetrical curves around the "Sen=Spe" line, depending on whether DOR is or is not constant. One method to test if DOR is constant, implemented by Meta-Disc in the threshold option, is the Moses-Shapiro-Littenberg model. With this model, testing the hypothesis of a constant DOR is equivalent to testing whether parameter b = 0. If P > 0.05 (indicate a constant DOR), which indicated that there was an absence of a notable threshold effect in the accuracy estimates among individual studies.

References
1. Devillé WL, Buntinx F, Bouter LM, Montori VM, de Vet HC, van der Windt DA, Bezemer PD.

Conducting systematic reviews of diagnostic studies: didactic guidelines. BMC Med Res Methodol 2002; 2:9.

VERSION 2 – REVIEW

REVIEWER	Steven Lane Department of Biostatistics University of Liverpool
REVIEW RETURNED	11-Sep-2015

GENERAL COMMENTS	The reviewer completed the checklist but made no further comments.