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)	Diagnostic Accuracy of Adrenal Imaging for Subtype Diagnosis in Primary Aldosteronism: Systematic Review and Meta-Analysis
AUTHORS	Zhou, Yaqiong; Wang, Dan; Jiang, Licheng; Ran, Fei; Chen, Sichao; Zhou, Peng; Wang, Peijian

VERSION 1 – REVIEW

REVIEWER	Qifu Li Chongqing medical university
REVIEW RETURNED	24-Mar-2020

GENERAL COMMENTS	In this manuscript, the authors evaluated the diagnostic accuracy of adrenal CT/MRI for the subtype diagnosis of PA using a method of meta-analysis. Unsurprisingly, CT/MRI is not a reliable alternative to AVS, which is widely recognized. However, this issue is important for clinical decision, and the systematic review and meta-analysis summarize new data from recent years (about 1/3 studies from recent 3 year). Here are some questions about this manuscript. 1. This study is a meta-analysis of diagnostic tests, while the reference test is not clearly described. It seems that AVS is the 'reference test' or golden criterion, which is mentioned in the part of 'Eligibility Criteria' (Page 6). Please provide a detailed description of the reference test, including the selectivity criterion, lateralization criterion, the number of patients diagnosed as unilateral right / unilateral left / bilateral / undefined side, the definition of successful AVS and technical AVS failure. All the diagnostic parameters seemed to be unreliable with uncertain reference test. 2. The authors included studies which performed AVS with or without ACTH stimulation. Usually, the reference test is an unified standard, while the authors provided two criteria of the reference test (with or without ACTH stimulation). 3. The cutoff value for the lateralization index (LI) in every included study (with or without ACTH stimulation) should be presented. If it is possible, please also provide the subgroup analysis based on different LI. 4. Why did the authors choose age, imaging methodology, publication date as the confounders and stratified by these parameters? Did the authors performed meta-regression? 5. For table 2, please provide the data of age> 40years. 6. The figure of Publication bias should be provided.

REVIEWER	John Funder Hudson Institute of Medical Research, and Monash University, Australia.
REVIEW RETURNED	04-Apr-2020

GENERAL COMMENTS	The manuscript by Zhou et al describes the concordance (or otherwise) of CT/MRI and AVS (as the accepted standard) in distinguishing unilateral from bilateral primary aldosteronism The comments I have a relatively minor, asa follows: 1. The last sentence of the Abstract, and the same sentence in the Conclusions, should be omitted - it gives hospitals an out, and allows them to settle for second best, to the ultimate detriment of the patient population. There are ways in which lateralization can be assisted in some patients by means other than CT/MRI in the absence of AVS - eg a very high level of 18-OH cortisol/18-oxocortisol, and similarly nomograms that point to BAH. 2. Page 4, lines 56 et ff. APA patients with biochemical cure after adrenalectomy have risk profiles slightly but significantly lower than age-, sex-, and BP-matched essential hypertensives. Patients with BAH in whom renin is not suppressed after 6 months of treatment have risk profiles indistinguishable from essential hypertensives; those in whom renin remains suppressed have risk profiles three times higher, which calls for additional intervention. It is not as simple nor as inevitable as the authors make out it to be. 3. Study Selection. The authors give reasons for reducing 60 to 25; a prior, much more sweeping reduction (1022 to 60) is presented without any reasons, which should be included and discussed. 4. The last paragraph before the limitations (page 14).should be omitted. There are so many flaws in the Spartacus study that it cannot not be taken as evidence. The last sentence of the paragraph is absolutely incorrect. 5. Just over half of the references cited are given as initials rather than names. This is novel, and I suspect should be corrected. 6. The English is often non-idiomatic/nongrammatical, and on one or two occions not east y to understand. in a revised version the final draft should be copy-edited by a native English-speaking scientist/doctor before submision.

REVIEWER	Naohiro Yonemoto NCNP, Japan
REVIEW RETURNED	23-Jul-2020

GENERAL COMMENTS	My comment is only from statistical view. 1. I recommend to share a table of original data by study in the meta-analysis. 2. You should describe the rationale of subgroup analysis. 3. You should add the funnel plot for checking the publication bias, not only the Deek test. 4. P8. L54, Why did you use cohen k test ? You should more clearly describe details of the data. If it was not matched, how do you do it ? 5. P8. L59 I think the setting in the test does not make sense. Delete it. 6. P10, L14, In table1, "Daisuke" ref 19 is not family name. 7.You should more describe the details of the mixed effect model for general readers. 8. The results find high heterogeneity. And then, you cannot directly interpret results of the overall analysis. The subgroup analysis would be important. You should more clearly discuss the results and limitations. 9. You should appropriately make up the digests in all tables

	10. P25, In the figure, "Sicheng 2019" and "Kupers 2012" might be incorrect,because the estimates and 95%CI were imbalance. You should check it.

VERSION 1 – AUTHOR RESPONSE

Reviewer # 1

1. This study is a meta-analysis of diagnostic tests, while the reference test is not clearly described. It seems that AVS is the 'reference test' or golden criterion, which is mentioned in the part of 'Eligibility Criteria' (Page 6). Please provide a detailed description of the reference test, including the selectivity criterion, lateralization criterion, the number of patients diagnosed as unilateral right / unilateral left / bilateral / undefined side, the definition of successful AVS and technical AVS failure. All the diagnostic parameters seemed to be unreliable with uncertain reference test.

Reply: We thank Reviewer 1 for this valuable comment. As commented by Reviewer 1, all the diagnostic parameters seem to be unreliable with uncertain reference test. Accordingly, eligibility criteria have been established. In our study, we included all studies used AVS as the standard reference. However, AVS procedure, with or without ACTH stimulation, is still a controversial debate and different cutoff values of SI and LI are used in different centers during the different AVS procedure[1]. Moreover, there is no wide consensus on the optimal cutoff for SI and LI currently[2]. Therefore, it is impossible to standardize more detailed AVS parameters when we select the literature. As the Reviewer 1 suggested, we have added more details about the AVS and original data in **Eligibility Criteria** section and **Table 1**.

2. The authors included studies which performed AVS with or without ACTH stimulation. Usually, the reference test is an unified standard, while the authors provided two criteria of the reference test (with or without ACTH stimulation).

Reply: We thank Reviewer 1 for this insightful comment. According to our eligibility criteria, we included all studies used AVS as the standard reference. However, AVS procedure, with or without ACTH stimulation, is still a controversial debate and there is no wide consensus and guideline on the topic[1]. Accordingly, it is impossible to standardize the criteria of the AVS (with or without ACTH stimulation) to ensure the comprehensiveness of the literature search. Therefore, we included a study if it used AVS procedure, with or without ACTH stimulation. However, to rule out the influence of different AVS procedure (with or without ACTH stimulation), we performed subgroup analysis stratified by different AVS procedure (with or without ACTH stimulation).

3. The cutoff value for the lateralization index (LI) in every included study (with or without ACTH stimulation) should be presented. If it is possible, please also provide the subgroup analysis based on different LI.

Reply: We thank Reviewer 1 for this insightful comment. In our revised manuscript, we have highlighted different cutoff values of LI in **Table 1** and provided the subgroup analysis based on different LI as "In a further stratified analysis by the LI demonstrated that specificity and specificity were higher when LI was≥4 versus LI was≥2 [69% (95% CI: 62 to 75 vs. 61% (95% CI: 37 to 80) and 59% (95% CI: 50 to 68) vs. 54% (95% CI: 46 to 75), respectively]. ***(Table 2)***" (**Subgroup analyses section**, paragraph 2).

4. Why did the authors choose age, imaging methodology, publication date as the confounders and stratified by these parameters? Did the authors performed meta-regression?

Reply: We thank Reviewer 1 for this insightful comment. CT and MRI scans are the most widely used imaging techniques. Several studies demonstrated that MRI has poorer resolution and slower acquisition, with risk of respiratory artifacts and is inferior to CT in PA subtype evaluation[3-6]. Contrast materials can improve the visibility of adrenal structures imaged by CT and MRI scans and might have a positive effect on diagnosis accuracy[7]. Thus, imaging methods and contrast materials were thought as confounders for subgroup analysis. Moreover, a large sample number may represent experienced interventional radiologists and the credibility of the included studies. Thus, sample size was thought as another confounder for subgroup analysis. Given the nonfunctioning adrenocortical adenoma (incidentaloma) rate is age-dependent, and recent studies indicate that, in PA, a typical unilateral nodule based on imaging before the age of 40 or 35 years is very probably an APA. Thus, age was thought as a confounder of the results. However, in our meta-analysis, the mean age of the identified studies ranged from 35 to 56, with only 1 study recruiting all participants aged<40 years. Therefore, in our revised manuscript, we moved the subgroup analysis stratified by age. In our revised manuscript, given that different cutoff for LI criteria and AVS procedure (with or without ACTH stimulation) might also affect the results of diagnosis accuracy[6], we also performed subgroup analysis stratified by these parameters. Thus in our revised manuscript, we chose imaging methodology (CT or CT/MRI), contrast used or not, AVS procedure (with or without ACTH stimulation), the cutoff value for LI (2 or 4) and sample size (divided by 100 subjects) as the confounders.

In our revised manuscript, we have described the rationale of subgroup analysis in **Statistical Analysis** section (paragraph 3) and added related results and comments in **Subgroup analyses** section (paragraph 2).

5. For table 2, please provide the data of age> 40years.

Reply: According to the reviewer's suggestion, the data of age> 40years has been added in the **Table 2**.

6. The figure of publication bias should be provided.

Reply: We thank reviewer 1 for this insightful comment. We have added Deeks' funnel plot for checking the publication bias, which is shown in Supplementary **Figure S3**.

Reviewer # 2

1. The last sentence of the Abstract, and the same sentence in the Conclusions, should be omitted - it gives hospitals an out, and allows them to settle for second best, to the ultimate detriment of the patient population. There are ways in which lateralization can be assisted in some patients by means other than CT/MRI in the absence of AVS - eg a very high level of 18-OH cortisol/18-oxocortisol, and similarly nomograms that point to BAH.

Reply: We thank Reviewer 2 for this constructive comment. We have removed these sentences. Moreover, we have added follow comments in **Discussion** sections (paragraph 6).

"If centers without AVS facilities currently, what should a physician do? The past few years have witnessed a rapidly growing interest in testing the utility of hybrid steroids, such as 18-oxocortisol/18-hydroxycortisol, for PA subtype and the results demonstrated that levels of 18-oxocortisol/18-hydroxycortisol plus an adenoma on CT/MRI might be of more assistance in those centers without AVS facilities especially in Japan and China, given their very high percentage of KCNJ5 mutations[8-10]. What will hopefully very substantially reduce or replace lateralization by AVS, perhaps the possibility of multi-steroid fingerprints in peripheral blood samples that distinguish unilateral from bilateral PA with a high degree of accuracy."

2. Page 4, lines 56 et ff. APA patients with biochemical cure after adrenalectomy have risk profiles slightly but significantly lower than age-, sex-, and BP-matched essential hypertensives. Patients with BAH in whom renin is not suppressed after 6 months of treatment have risk profiles indistinguishable from essential hypertensives; those in whom renin remains suppressed have risk profiles three times higher, which calls for additional intervention. It is not as simple nor as inevitable as the authors make out it to be.

Reply: We thank Reviewer 2 for this comment. We have made revisions as *"Accumulating clinical and epidemiological evidence suggests that PA amplifies cardiovascular and cerebrovascular complications beyond essential hypertension prior to treatment, even after controlling for the elevated blood pressure. However, patients with the unilateral resected PA and post-adrenalectomy biochemically cured patients had slightly better risk profiles than matched essential hypertensive patients. Patients with bilateral PA whose plasma renin activity is not suppressed after 6 months of* *mineralocorticoid receptor antagonists (MRA) therapy have the same risk profiles as essential hypertensive patients; those whose renin activity remains suppressed have 4-fold higher risk profiles than controls and titration of MRA therapy to raise renin might reduce this excess risk*[11,12]*"* (**Introduction** section, paragraph 1)

3. Study Selection. The authors give reasons for reducing 60 to 25; a prior, much more sweeping reduction (1022 to 60) is presented without any reasons, which should be included and discussed.

Reply: We thank Reviewer 2 for this comment. We have listed reasons for the sweeping articles exclusion and the related comments were added as *"Among them 962 studies excluded for the following reasons: 489 studies were not relevant; 280 studies were reviews or practice guidelines; 92 studies did not include humans; 101studies were case/letter report"* (**Study Selection** section, paragraph 1).

4. The last paragraph before the limitations (page 14) should be omitted. There are so many flaws in the Spartacus study that it cannot be taken as evidence. The last sentence of the paragraph is absolutely incorrect.

Reply: We thank Reviewer 2 for this comment. After careful consideration, we find that there are several caveats to generalization of Spartacus conclusions[13]. They selected the most severe PA patients and the difficulties in reconciling CT diagnoses between cooperating centers, evidenced by a notable difference between right versus left adenomas on CT and AVS, are unsettling. Therefore, as suggested by Reviewer 2, we have dropped these sentences.

5. Just over half of the references cited are given as initials rather than names. This is novel, and I suspect should be corrected.

Reply: We thank the careful review of Reviewer 2. We have revised the references carefully according to the requirements of *BMJ Open*.

6. The English is often non-idiomatic/nongrammatical, and on one or two occsons not easy to understand. In a revised version the final draft should be copy-edited by a native English-speaking scientist/doctor before submission.

Reply: We thank Reviewer 2 for this comment. Our revised manuscript have been copy-edited by a native English-speaking doctor

Reviewer# 3

1. I recommend sharing a table of original data by study in the meta-analysis.

Reply: We thank Reviewer 3 for this comment. We have added original data in **Table 1.**

2. You should describe the rationale of subgroup analysis.

Reply: We thank Reviewer 3 for this insightful comment. CT and MRI scans are the most widely used imaging techniques. Several studies demonstrated that MRI has poorer resolution and slower acquisition, with risk of respiratory artifacts and is inferior to CT in PA subtype evaluation[3-6]. Contrast materials can improve the visibility of adrenal structures imaged by CT and MRI scans and might have a positive effect on diagnosis accuracy[7]. Thus, imaging methods and contrast materials were thought as confounders for subgroup analysis. Moreover, a large sample number may represent experienced interventional radiologists and the credibility of the included studies. Thus, sample size was thought as another confounder for subgroup analysis. Given the nonfunctioning adrenocortical adenoma (incidentaloma) rate is age-dependent, and recent studies indicate that, in PA, a typical unilateral nodule based on imaging before the age of 40 or 35 years is very probably an APA. Thus, age was thought as a confounder of the results. However, in our meta-analysis, the mean age of the identified studies ranged from 35 to 56, with only 1 study recruiting all participants aged<40 years. Therefore, in our revised manuscript, we moved the subgroup analysis stratified by age. In our revised manuscript, given that different cutoff for LI criteria and AVS procedure (with or without ACTH stimulation) might also affect the results of diagnosis accuracy[6], we also performed subgroup analysis stratified by these parameters. Thus in our revised manuscript, we chose imaging methodology (CT or CT/MRI), contrast used or not, AVS procedure (with or without ACTH stimulation), the cutoff value for LI (2 or 4) and sample size (divided by 100 subjects) as the confounders.

In our revised manuscript, we have described the rationale of subgroup analysis in **Statistical Analysis** section (paragraph 3) and added related results and comments in **Subgroup analyses** section (paragraph 2).

3. You should add the funnel plot for checking the publication bias, not only the Deek test.

Reply: We have added Deeks' funnel plot for checking the publication bias, which is shown in Supplementary **Figure S3**.

4. P8. L54,Why did you use cohen k test? You should more clearly describe details of the data. If it was not matched, how do you do it?

Reply: We thank Reviewer 3 for this comment. The methodological quality of identified studies was assessed by 2 independent reviewers using the QUADAS-2 criteria and the results were summarized in **Table S2**. If it was not matched, a third reviewer was involved for disagreements and final decisions were determined by consensus. We have described details of the data in **Quality Assessment** section as "*Inter-rater agreement for QUADAS-2 results was assessed by using Cohen's κ coefficient. Out of 175 QUADAS-2 items (25 articles×7 items), the 2 reviewers agreed on 172 (98%) with an inter-rater agreement of κ=0.9*". Moreover, the quality assessment for each study by 2 reviewers is shown in **Table S3.**

5. P8. L59 I think the setting in the test does not make sense. Delete it.

Reply: We thank Reviewer 3 for this comment. We have deleted the setting in the test.

6. P10, L14, In table1, "Daisuke" ref 19 is not family name.

Reply: We thank the careful review of Reviewer 3. We have revised the references carefully according to the requirements of *BMJ Open*.

7. You should more describe the details of the mixed effect model for general readers.

Reply: We thank Reviewer 3 for this comment. We have more described the details of the mixed effect model as "*This approach assumes bivariate normal distributions for the logit transformations of sensitivity and specificity from individual studies. The parameters of the bivariate model are estimated in a single model to incorporate the possible correlation between sensitivities and specificities. These bivariate models can be analyzed using linear mixed model techniques that are now widely available in statistical packages, such as STATA gllamm* [14,15]" in the **Statistical Analysis** section (paragraph 1).

8. The results find high heterogeneity. And then, you cannot directly interpret results of the overall analysis. The subgroup analysis would be important. You should more clearly discuss the results and limitations.

Reply: We thank Reviewer 3 for this insightful comment. In our study, heterogeneity is a potential problem that may affect the interpretation of the results. Totally speaking, diversity in design, sample-size, inclusion criteria, complex population composition and other unknown reasons all can lead to the heterogeneity. In our revised manuscript, subgroup and meta-regression analyses were performed to explore the possible sources of heterogeneity. Moreover, we have used random effects model to minimize the influence of heterogeneity on our results[16].

As commented by Reviewer 3 that we cannot directly interpret results of the overall analysis and the subgroup analysis would be important. As suggested by reviewer 3, in our revised manuscript, we firstly concluded that "CT/MRI has a poor sensitivity (68 %) and specificity (57 %) in the subtype classification when used AVS as the reference standard for unilateral PA" in the second paragraph of Discussion section. Subsequently, we more clearly discussed the results of subgroup analyses based on contrast-enhanced CT, AVS procedure (with or without ACTH stimulation), age and the cutoff of LI. In the end, we have explained the heterogeneity in the **Limitation** section.

9. You should appropriately make up the digests in all tables

Reply: We have made up the digests in all tables.

10. P25. In the figure, "Sicheng 2019" and "Kupers 2012" might be incorrect, because the estimates and 95%CI were imbalance. You should check it.

Reply: We thank Reviewer 3 for this comment. After examining the data of these two included studies carefully by 2 independent investigators (Z.Y.Q and W.P.J), we found that the data of 'Kűpers 2012' was incorrect. In the previous version, the true positive, true negative, false positive, and false negative were 27, 5, 22, and 33, respectively. In the revised version, the true positive, true negative, false positive, and false negative were 26, 5, 23, and 33, respectively. The figures and tables were also revised. There was no error in "Sicheng 2019". The revised analysis did not alter the conclusion of the study

References:

1. I L, M A, F Z,et.al. Adrenal venous sampling with or without adrenocorticotropic hormone stimulation: A meta-analysis. 2018.
2. GP R, G M, TM S. Adrenal Venous Sampling: Where Do We Stand? , 2019:843-858.
3. Asmar M, Wachtel H, Yan Y,et.al. Reversing the established order: Should adrenal venous sampling precede cross-sectional imaging in the evaluation of primary aldosteronism? *Journal of surgical oncology* 2015;112:144-148
4. Pedersen M, Karlsen MA, Ankjærgaard KL,et.al. Primary hyperaldosteronism diagnosed with adrenal vein sampling. Characteristics and follow-up after adrenalectomy in a Danish study. *Scandinavian journal of clinical and laboratory investigation* 2016;76:45-50
5. Campbell RA, Young DS, Shaver CN,et.al. Influence of Adrenal Venous Sampling on Management in Patients with Primary Aldosteronism Independent of Lateralization on Cross-Sectional Imaging. *Journal of the American College of Surgeons* 2019;229:116-124
6. Sam D, Kline GA, So B,et.al. Discordance Between Imaging and Adrenal Vein Sampling in Primary Aldosteronism Irrespective of Interpretation Criteria. *The Journal of clinical endocrinology and metabolism* 2019;104:1900-1906
7. Nandra G, Duxbury O, Patel P,et.al. Technical and Interpretive Pitfalls in Adrenal Imaging. 2020:1041-1060.
8. Taguchi R, Yamada M, Nakajima Y,et.al. Expression and mutations of KCNJ5 mRNA in Japanese

patients with aldosterone-producing adenomas. 2012:1311-1319.
9. Wang B, Li X, Zhang X,et.al. Prevalence and characterization of somatic mutations in Chinese aldosterone-producing adenoma patients. 2015:e708.
10. F S, R M, Y O,et.al. Measurement of peripheral plasma 18-oxocortisol can discriminate unilateral adenoma from bilateral diseases in patients with primary aldosteronism. 2015:1096-1102.
11. Hundemer GL, Curhan GC, Yozamp N,et.al. Cardiometabolic outcomes and mortality in medically treated primary aldosteronism: a retrospective cohort study. 2018:51-59.
12. Funder JW. Primary Aldosteronism: Where Are We Now? Where to from Here? , 2020:459-466.
13. Dekkers T, Prejbisz A, Kool LJS,et.al. Adrenal vein sampling versus CT scan to determine treatment in primary aldosteronism: an outcome-based randomised diagnostic trial. 2016:739-746.
14. Reitsma JB, Glas AS, Rutjes AW,et.al. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. 2005:982-990.
15. H C, H G, Y Z. Bivariate random effects meta-analysis of diagnostic studies using generalized linear mixed models. 2010:499-508.
16. H C, H G, Y Z. Bivariate random effects meta-analysis of diagnostic studies using generalized linear mixed models. *Medical decision making : an international journal of the Society for Medical Decision Making* 2010;30:499-508.PMID:19959794

VERSION 2 – REVIEW

REVIEWER	Qifu Li Ist affiliated hospital of Chonhqing Medical University. P. R. China
REVIEW RETURNED	11-Sep-2020

GENERAL COMMENTS	This study was well conducted based on rigorous statistical methods, but there were still some issues that should be addressed as follows: Major issues 1. In the Fig.3, the I2 of sensitivity and specificity were 86.89% and 86.88% which meant fiercely high heterogeneity, may leading the unreliable results. Please explain it in the discussion part and may add some subgroup analysis to tailor precise diagnosis strategies. 2. A subgroup analysis stratified by the follow-up data should be performed ,which can confirm that adrenalectomies were performed appropriately. Minor issues 1. In Search Strategy, it is said that Trials in abstract form without a published manuscript also excluded. It is recognized that the unpublished data showed much lower level of evidence (LOE) and was searched in different way from published articles. Please clarify the search process of abstract form without a published manuscript and the search outcomes. 2. In the Fig 1, 35 articles were excluded with reasons after full review. Please stated the detailed reasons for excluding, since even one article could lead to the opposite conclusion. 3. Whether there are detailed diagnostic criteria of PA in studies is a confounding factor affecting the subtype classification , which should be included in the heterogeneity analysis. Similarly, age was thought as confounders for adrenal CT in PA subtype classification, which should also be included in the heterogeneity analysis.

REVIEWER	John Funder

	Hudson Institute of Medical Research and Monash University, Australia
REVIEW RETURNED	09-Sep-2020

GENERAL COMMENTS	The manuscript needs major editing for grammar, word choice, idiom and usage before it can be published. If this is done at the BMJ, so be it. If this needs to be done elswhere, the authors need to engage the services of a native English speaking colleague - not a translation service - to bring it to an acceptable standard.

REVIEWER	Naohiro Yonemoto NCNP, Japan
REVIEW RETURNED	03-Sep-2020

GENERAL COMMENTS	Thank you for revised manuscript. Unfortunately, I cannot find the improvement for it. I think the analysis in overall was not sufficient for the heterogeneity. The description of data should be needed for the replication. Also, I cannot find the contribution of clinical perspective in the meta-analysis.

VERSION 2 – AUTHOR RESPONSE

Reviewer # 1

Comments 1: I think the analysis in overall was not sufficient for the heterogeneity.

Response: We thank Reviewer 1 for this comment. As commented by Reviewer 1 that the analysis in overall was not sufficient for the heterogeneity, the subgroup analysis would be important. In this revised manuscript, we further performed subgroup analyses based on the diagnostic criteria of PA and methodological quality. In addition, meta-regression and sensitivity analysis were also performed to explore the possible sources of heterogeneity. Moreover, we have used random-effects model to minimize the influence of heterogeneity on our results.

Subgroup analysis stratified by screening test showed that the sensitivity of the ARR group was higher than that of the non-ARR group (78% vs. 66%). The heterogeneity was high (I^2 =87.7%) in the ARR subgroup, whereas it disappeared (0%) in the non-ARR subgroup. Regarding specificity, the heterogeneity was high for both groups (86.2% vs. 89.1%). Subgroup analysis stratified by the confirmatory test for PA demonstrated an increase in sensitivity (71% vs. 57%) and a slight decrease in specificity (60% vs. 66%) for the salt-loading test group compared with additional options group, with significant heterogeneity observed in all the above groups (I^2 all > 50%).

Subgroup analysis based on methodological quality (high-quality, low-quality and unclear-quality) revealed that there was low heterogeneity for sensitivity in all the above groups (I^2 all< 50%). The diagnostic pooled sensitivity for the high-quality group was the highest, followed by the unclear-quality group and the low-quality group (78% vs. 62% vs. 48%). The unclear-quality group had the highest specificity, followed by the high-quality group and the low-quality group (69% vs. 62% vs. 51%). Regarding specificity, heterogeneity was significantly decreased but still high in all the groups.

Results of meta-regression analysis showed that the sample size was the only covariate with a negative effect on sensitivity. Additionally, there was a significant interaction between lower age, as well as high methodological quality, and higher specificity of CT/MRI for the detection of unilateral forms of PA (Figure S2), which were detailed in **Meta-regression analysis** section.

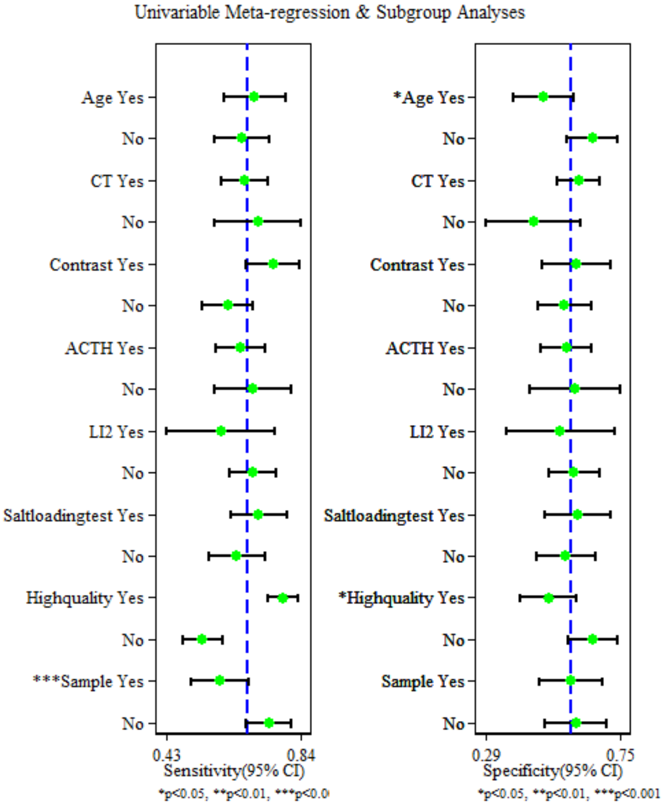


Figure S2: Graphical presentation of the generalized linear mixed model exploring the impact of selected variables on sensitivity and specificity.

Goodness-of-fit and bivariate normality analyses (**Figures S3a, S3b**) showed that the bivariate model was moderately robust. Influence analysis and outlier detection identified 4 outliers (**Figures S3c, S3d**). After we excluded these outliers, the overall results did not change significantly, which suggested that the results of this study were statistically reliable (**Table 2**).The above comments were added in **Sensitivity analysis** section.

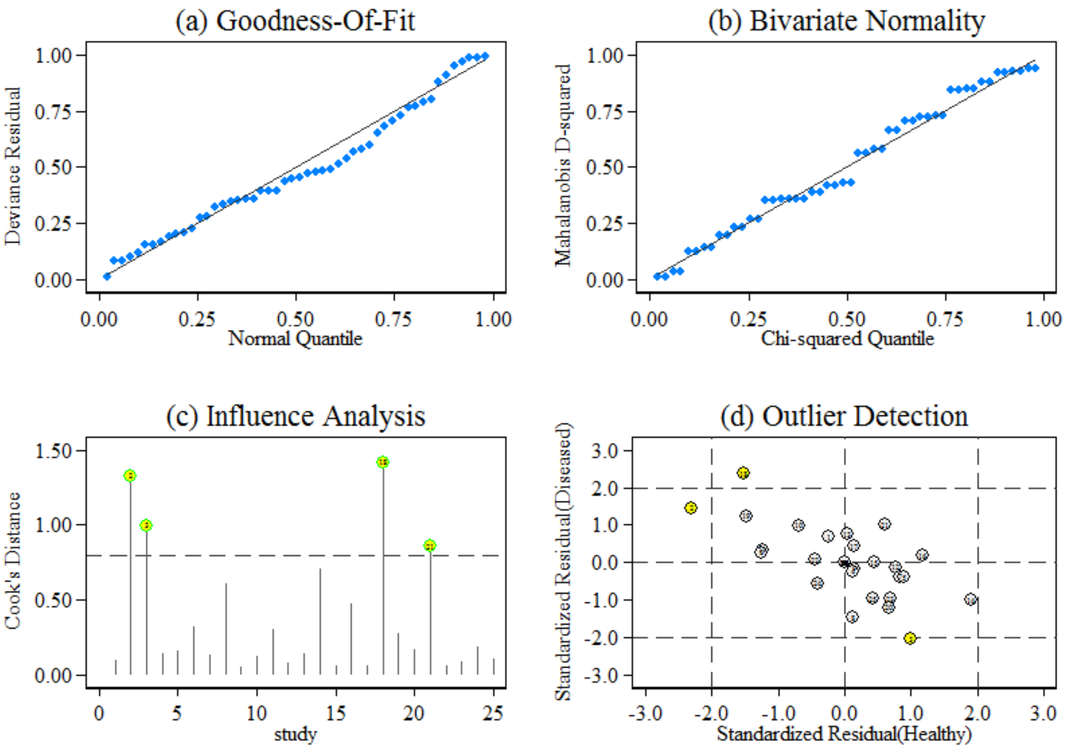


Figure S3 Graphs for sensitivity analyses: a goodness of fit, b bivariate normality, c influence analysis, and d outlier detection.

Moreover, the follow comments were added in **Discussion** section.

The present meta-analysis has several limitations. First, there was great heterogeneity among the included studies, which might have compromised the credibility. The results of the subgroup analyses and meta-regression suggested that the screening test for PA, age, study quality, sample size, and other unknown factors may also contribute to the aforementioned heterogeneity. However, the results from the subgroup analyses and sensitivity analysis all confirmed the robustness of our meta-analysis’s results

Comments 2: The description of data should be needed for the replication.

Response: We thank Reviewer 1 for this comment. Actually, we have added original data in **Table 1** and **Table S3**. Moreover, we have described the rationale of subgroup analyses and made a detailed data description in **Statistical Analysis** section**.** Based on these data, the results can be repeated easily.

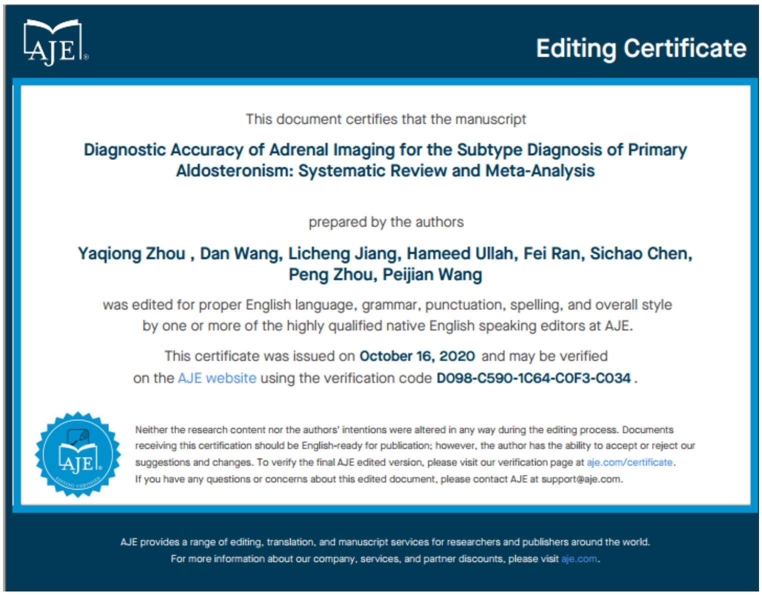
Comments 3: Also, I cannot find the contribution of clinical perspective in the meta-analysis.

Response: We thank Reviewer 1 for this comment. The contribution of clinical perspective in the meta-analysis was mainly featured in following aspects: 1) to our best knowledge, this is the first meta-analysis to investigate the diagnostic accuracy of CT/MRI for the subtype diagnosis of PA, involving 4669 individuals from 25 studies; 2) many physicians prefer to perform CT/MRI as the first and sometimes the only investigation of subtype diagnosis. However, according to the present meta-analysis, CT/MRI has poor sensitivity and specificity in the detection of unilateral PA; 3) the 2009 guidelines for managing PA contended that younger patients(≤40 years) with an unequivocal biochemical diagnosis of PA and a clear-cut unilateral adenoma on adrenal CT scan proceed directly to surgery and AVS procedure may be skipped. However, based on this meta-analysis, even in young patients (≤40 years), 21% would have undergone unnecessary surgery based on imaging results alone. Therefore, we recommend routinely referring all patients for AVS, regardless of age and imaging results. In our revised manuscript, we have described the details of the contribution of clinical perspective in the **Discussion** section.

Reviewer# 2

Comments 1.The manuscript needs major editing for grammar, word choice, idiom and usage before it can be published. If this is done at the BMJ, so be it. If this needs to be done elsewhere, the authors need to engage the services of a native English speaking colleague - not a translation service - to bring it to an acceptable standard.

Response: We thank Reviewer 2 for this comment. As suggested by Reviewer 2, our manuscript has been language edited by American Journal Experts. The editing certificate was show in **Figure S5**.



AJE®

Editing Certificate

This document certifies that the manuscript

Diagnostic Accuracy of Adrenal Imaging for the Subtype Diagnosis of Primary Aldosteronism: Systematic Review and Meta-Analysis

prepared by the authors

Yaqiong Zhou , Dan Wang, Licheng Jiang, Hameed Ullah, Fei Ran, Sichao Chen, Peng Zhou, Peijian Wang

was edited for proper English language, grammar, punctuation, spelling, and overall style by one or more of the highly qualified native English speaking editors at AJE.

This certificate was issued on **October 16, 2020** and may be verified on the AJE website using the verification code **D098-C590-1C64-C0F3-C034** .

Neither the research content nor the authors' intentions were altered in any way during the editing process. Documents receiving this certification should be English-ready for publication; however, the author has the ability to accept or reject our suggestions and changes. To verify the final AJE edited version, please visit our verification page at aje.com/certificate. If you have any questions or concerns about this edited document, please contact AJE at support@aje.com.

AJE provides a range of editing, translation, and manuscript services for researchers and publishers around the world. For more information about our company, services, and partner discounts, please visit aje.com.

Figure S5: The editing certificate

Reviewer# 3

Major issues

Comments 1.In the Fig.3, the I2 of sensitivity and specificity were 86.89% and 86.88% which meant fiercely high heterogeneity, may leading the unreliable results. Please explain it in the discussion part and may add some subgroup analysis to tailor precise diagnosis strategies.

Response: We thank Reviewer 3 for this insightful comment. As commented by Reviewer 3 that the I^2 of sensitivity and specificity were 86.89% and 86.88% which meant fiercely high heterogeneity, may leading the unreliable results. In this revised manuscript, we further performed subgroup analyses based on the diagnostic criteria of PA and methodological quality. In addition, meta-regression and sensitivity analysis were also performed to explore the possible sources of heterogeneity. Moreover, we have used random effects model to minimize the influence of heterogeneity on our results.

Subgroup analysis stratified by screening test showed that the sensitivity of the ARR group was higher than that of the non-ARR group (78% vs. 66%). The heterogeneity was high (I^2 =87.7%) in the ARR subgroup, whereas it disappeared (0%) in the non-ARR subgroup. Regarding specificity, the heterogeneity was high for both groups (86.2% vs. 89.1%). Subgroup analysis stratified by the confirmatory test for PA demonstrated an increase in sensitivity (71% vs. 57%) and a slight decrease in specificity (60% vs. 66%) for the salt-loading test group compared with additional options group, with significant heterogeneity observed in all the above groups (I^2 all > 50%).

Subgroup analysis based on methodological quality (high-quality, low-quality and unclear-quality) revealed that there was low heterogeneity for sensitivity in all the above groups (I^2 all< 50%). The diagnostic pooled sensitivity for the high-quality group was the highest, followed by the unclear-quality group and the low-quality group (78% vs. 62% vs. 48%). The unclear-quality group had the highest specificity, followed by the high-quality group and the low-quality group (69% vs. 62% vs. 51%). Regarding specificity, heterogeneity was significantly decreased but still high in all the groups.

Results of meta-regression analysis showed that the sample size was the only covariate with a negative effect on sensitivity. Additionally, there was a significant interaction between lower age, as well as high methodological quality, and higher specificity of CT/MRI for the detection of unilateral forms of PA (**Figure S2**), which were detailed in **Meta-regression analysis** section.

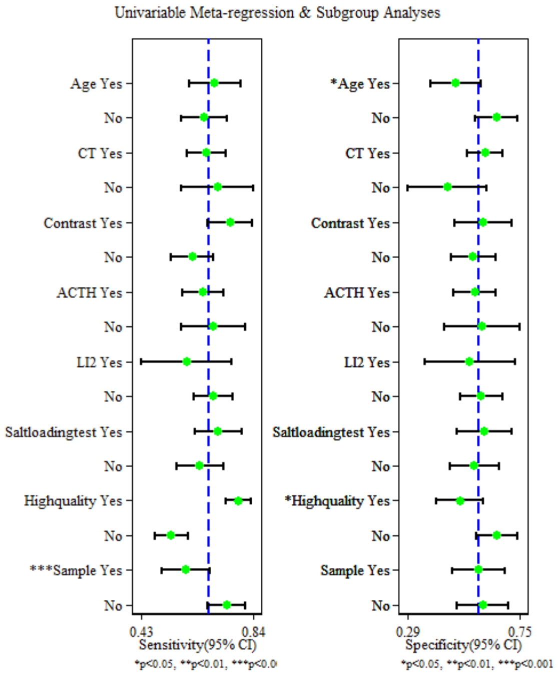


Figure S2: Graphical presentation of the generalized linear mixed model exploring the impact of selected variables on sensitivity and specificity.

Goodness-of-fit and bivariate normality analyses (Figures S3a, S3b) showed that the bivariate model was moderately robust. Influence analysis and outlier detection identified 4 outliers (Figures S3c, S3d). After we excluded these outliers, the overall results did not change significantly, which suggested that the results of this study were statistically reliable (**Table 2**).The above comments were added in **Sensitivity analysis** section.

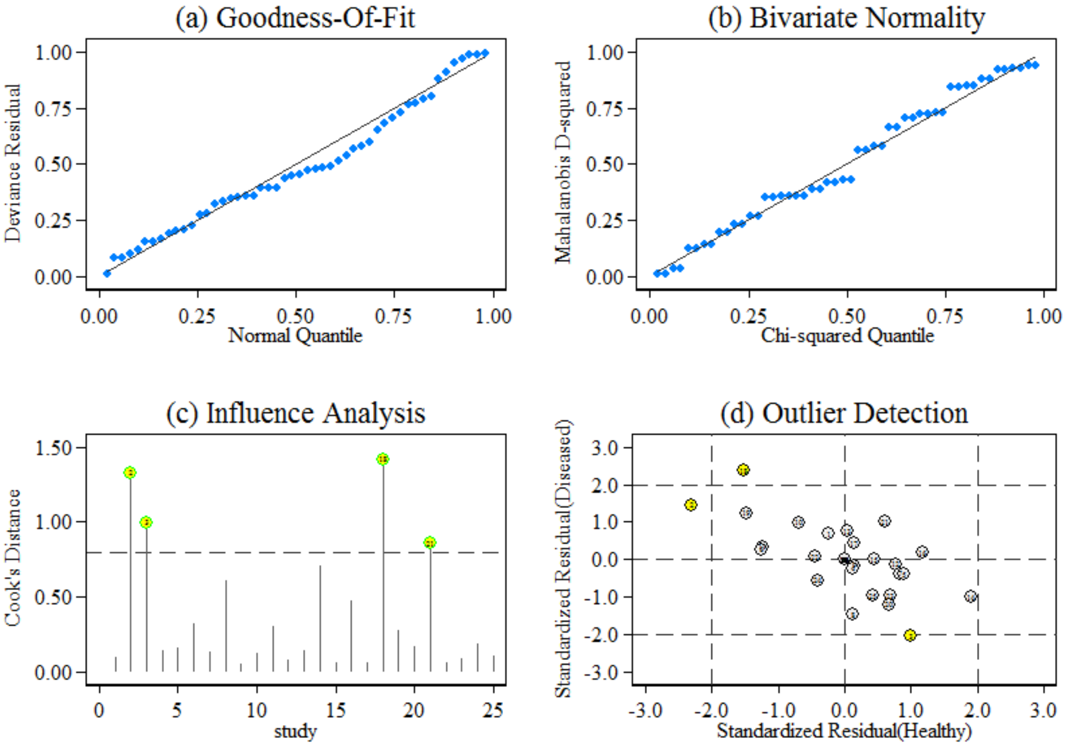


Figure S3 Graphs for sensitivity analyses: a goodness of fit, b bivariate normality, c influence analysis, and d outlier detection.

Moreover, the follow comments were added in **Discussion** section.

The present meta-analysis has several limitations. First, there was great heterogeneity among the included studies, which might have compromised the credibility. The results of the subgroup analyses and meta-regression suggested that the screening test for PA, age, study quality, sample size, and other unknown factors may also contribute to the aforementioned heterogeneity. However, the results from the subgroup analyses and sensitivity analysis all confirmed the robustness of our meta-analysis's results

Comments 2.A subgroup analysis stratified by the follow-up data should be performed, which can confirm that adrenalectomies were performed appropriately.

Response: We thank Reviewer 3 for this insightful comment. Among studies included in the present meta-analysis, there were 10 studies reported surgical outcomes for unilateral PA. In these studies, only 2 articles compared clinical outcomes after adrenalectomy between AVS-guide and CT-guide. The remaining 8 studies did not compare the clinical outcomes between the two groups. Although pathologic data were available in the 8 studies, information on biochemical measures, blood pressure, and medication use during follow-up was too sparse. Moreover, the lack of unified criteria of evaluation the effect of surgery and the comparability of the included studies, limited our ability to perform pooled analysis. Therefore, based on these data, a subgroup analysis stratified by the follow-up data cannot be performed.

Minor issues

Comments 1. In Search Strategy, it is said that Trials in abstract form without a published manuscript also excluded. It is recognized that the unpublished data showed much lower level of evidence (LOE) and was searched in different way from published articles. Please clarify the search process of abstract form without a published manuscript and the search outcomes.

Response: We thank Reviewer 3 for this comment. We originally intended to express that trials in abstract form without **a full text** were also excluded. To avoid confusion, we removed the sentence “Trials in abstract form without a published manuscript also excluded”.

Comments 2. In the Fig 1, 35 articles were excluded with reasons after full review. Please state the detailed reasons for excluding, since even one article could lead to the opposite conclusion.

Response: We thank Reviewer 3 for this comment. A total of 35 studies were excluded for the following reasons: data to compute diagnostic accuracy were not provided or could not be derived (25 papers), reporting on the same population (4 papers) and no comparison of CT/MRI and AVS results in individual patients (6 papers). In the revised manuscript, we have added the above information in the **Figure 1**.

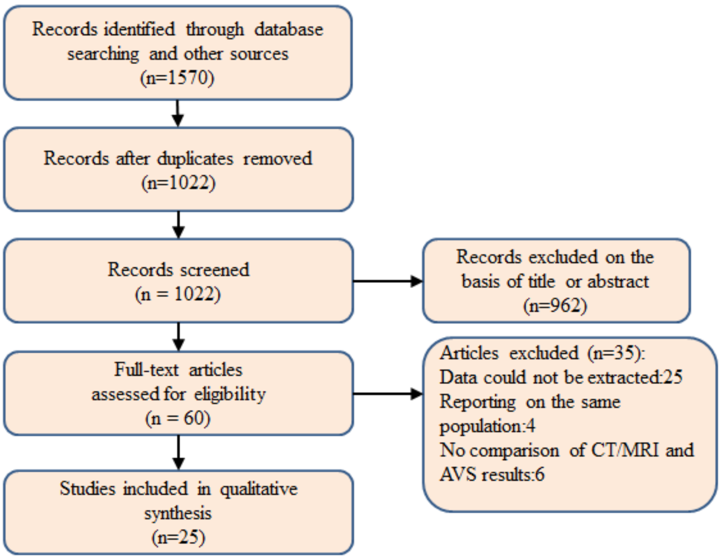

Records identified through database searching and other sources (n=1570)
Records after duplicates removed (n=1022)
Records screened (n = 1022)
Records excluded on the basis of title or abstract (n=962)
Full-text articles assessed for eligibility (n = 60)
Articles excluded (n=35):
Data could not be extracted:25
Reporting on the same population:4
No comparison of CT/MRI and AVS results:6
Studies included in qualitative synthesis (n=25)

Figure 1: Flow diagram of the review process.

Comments 3.Whether there are detailed diagnostic criteria of PA in studies is a confounding factor affecting the subtype classification, which should be included in the heterogeneity analysis. Similarly, age was thought as confounders for adrenal CT in PA subtype classification, which should also be included in the heterogeneity analysis.

Response: We thank Reviewer 3 for this comment. It is true that the diagnostic criteria of PA and age may significantly influence diagnostic accuracy. Therefore, as suggested by Reviewer 3, we investigated the impact of these parameters on the heterogeneity by meta-regression analysis.

In our revised manuscript, we ultimately chose imaging methodology, contrast used or not, AVS procedure, the cutoff value for LI, sample size, age, confirmatory testing for PA and quality of studies as the confounders. Results of meta-regression analysis showed that the sample size was the only covariate with a negative effect on sensitivity. Additionally, there was a significant interaction between lower age, as well as high methodological quality, and higher specificity of CT/MRI for the detection of unilateral forms of PA (**Figure S2**), which were detailed in **Meta-regression analysis** section.

VERSION 3 – REVIEW

REVIEWER	Qifu Li Chongqing Medical university P.R.China
REVIEW RETURNED	04-Nov-2020

GENERAL COMMENTS	All questions were well answered!

REVIEWER	Naohiro Yonemoto NCNP, Japan
REVIEW RETURNED	23-Oct-2020

GENERAL COMMENTS	Thank you for the revised. but I have still major concerns. I have some comments below. 1. The results are not reflected in the conclusions, despite you changed and added the results. 2. Overall analysis is useless due to high heterogeneity . You should only use them from subgroup. 3. I am not sure the word of "significant" was appropriate. 4. I am not clear what is the main hypothesis in the analysis. Can you test it from the data. I worry about the data is really relevant to it ? 5. This is major concerns. I cannot find revised letter on one by one comments. I cannot completely understand how did you think the reviewer comments and how to revise them or not.

VERSION 3 – AUTHOR RESPONSE

Reviewer #3

Comments 1: The results are not reflected in the conclusions, despite you changed and added the results.

Response: We thank Reviewer 3 for this comment. The main results of this meta-analysis have been reflected in the following aspects:
1) The overall analysis showed that CT/MRI has poor sensitivity (68%) and specificity (57%) in subtype classification when AVS was used as the reference standard. We have elaborated on this point in Discussion sections (paragraph 1, 2);
2) Subgroup analysis based on AVS procedure (with or without ACTH stimulation) revealed that there was no significant difference between the two AVS procedures and stricter thresholds for determining lateralization on AVS would result in higher sensitivity and specificity. We more clearly discussed the results of subgroup analyses based on AVS procedure and cut-off value for the LI in Discussion sections (paragraph 3);
3) Subgroup analysis stratified by screening test for PA showed that the sensitivity of the ARR group was higher than that of the non-ARR group (78% vs. 66%). Subgroup analysis stratified by the confirmatory test for PA demonstrated an increase in sensitivity (71% vs. 57%) and a slight decrease in specificity (60% vs. 66%) for the salt-loading test group compared with additional options group. We more clearly discussed the results of subgroup analyses based on screening test and confirmatory test for PA in Discussion sections (paragraph 4);
4) Subgroup analysis demonstrated that although the sensitivity (71%) and specificity (79%) were improved in young people (≤40 years). We more clearly discussed the results in Discussion sections (paragraph 5).
Based on the results, we conclude that CT/MRI has poor sensitivity (68%) and specificity (57%) in the detection of unilateral PA when AVS is used as the reference standard. Even in young patients (≤40 years), 21% would have undergone unnecessary adrenalectomy based on imaging results alone. The conclusion reflected the main findings of the results appropriately with great value of clinical perspective.

Comments 2: Overall analysis is useless due to high heterogeneity. You should only use them from subgroup.
Response: We thank Reviewer 3 for this comment. In theory, heterogeneity is inevitable in any meta-analyses. If there was high heterogeneity, what should authors do? On the one hand, random-effects model can be applied in methodology; on the other hand, it is essential to explore sources of heterogeneity and to incorporate it explicitly in the analysis. As such, heterogeneity offers opportunities for increasing our knowledge, rather than threats to our efforts to synthesize the available evidence. In our manuscript, we also adopted these two strategies.
Moreover, the objective of meta-analyses is to provide a summary estimate of enhanced precision from a series of diagnostic test evaluations, even with high heterogeneity. We believe that all results should be presented objectively to the readers rather than selectively reporting, regardless of heterogeneity. In our previous manuscript, we made a detailed analysis of the heterogeneity and informed the readers that our study should be interpreted with caution due to significant heterogeneity. Therefore, the overall analysis still has value even though there is high heterogeneity. However, as commented previously that moderate to high heterogeneity among the included studies might have compromised the credibility, which has been addressed in our previous revised manuscript.

Comments 3: I am not sure the word of "significant" was appropriate.
Response: We thank Reviewer 3 for this comment. After careful search, we found that the word of "significant" was used in six places in the whole text.
1) The sentence “Meta-regression revealed a significant impact of sample size on sensitivity and of age and study quality on specificity” in Abstract section (paragraph 3). We think that the word of "significant" was appropriate in this sentence.
2) The sentence “Subgroup analysis stratified by the confirmatory test for PA demonstrated an increase in sensitivity (71% vs. 57%) and a slight decrease in specificity (60% vs. 66%) for the salt-loading test group compared with additional options group, with significant heterogeneity observed in

all the above groups (I2 all > 50%)" in Subgroup analyses section (paragraph 3) . After careful consideration, we think that the word of "moderate to high heterogeneity" might be more appropriate in this sentence. Corrections have been made in the revised version.
3) The sentence "Additionally, there was a significant interaction between lower age, as well as high methodological quality, and higher specificity of CT/MRI for the detection of unilateral forms of PA" in Meta-regression analysis section (paragraph 1). We think that the word of "significant" was appropriate in this sentence.
4) The sentence "The present meta-analysis revealed that there was no significant difference between the two AVS procedures (with or without ACTH stimulation)" in Discussion section (paragraph 3). We think that the word of "significant" was appropriate in this sentence.
5) The sentence "the results should be interpreted with caution because of significant heterogeneity due to several underlying confounders" in Discussion section (paragraph 4). After careful consideration, we think that the word of "moderate to high heterogeneity" might be more appropriate in this sentence. Corrections have been made in the revised version.
6) The sentence "However, due to significant heterogeneity, our study should be interpreted with caution" in Conclusion section (paragraph 1). After careful consideration, we think that the word of "moderate to high heterogeneity" might be more appropriate in this sentence. Corrections have been made in the revised version.

Comments 4: I am not clear what is the main hypothesis in the analysis. Can you test it from the data? I worry about the data is really relevant to it?
Response: We thank Reviewer 3 for this comment. We have described main hypothesis in the analysis in the Abstract section (paragraph 1), Introduction section (paragraph 3) and Discussion sections (paragraph 1,2). The research of our proposition was launched around this hypothesis and purpose. Accurate subtype classification in PA is critical in assessing the optimal treatment options. Owing to less invasive nature, lower cost and wide availability, many physicians prefer to perform CT/MRI as the first, and sometimes only, investigation of PA subtype. By now, numerous studies have evaluated the diagnostic performance of CT/MRI in subtype diagnosis of PA, but the results have been inconsistent. Moreover, all these studies were limited by small sample sizes in a single centre, which limited the credibility of the results. In this context, we thus performed a comprehensive meta-analysis of all the available studies to evaluate the diagnostic value of CT/MRI for subtype classification of PA. This is the main hypothesis and purpose of this study.
Based on our results, we concluded that CT/MRI is not a reliable alternative to invasive AVS without excellent sensitivity or specificity for correctly identifying unilateral PA. Even in young patients (≤40 years), 21% of patients would have undergone unnecessary adrenalectomy based on imaging results alone.

Comments 5: This is major concerns. I cannot find revised letter on one by one comments. I cannot completely understand how did you think the reviewer comments and how to revise them or not.
Response: We thank Reviewer 3 for this comment. Actually, we uploaded the point-by-point author response letter when we uploaded the previous revised draft. Moreover, we also marked the changes within the document by using coloured text to help reviewers review the manuscript easily.
Thank you and best regards.