Author's response to reviews

Title: Does Risk for Ovarian Malignancy Algorithm excel human epididymis protein 4 and CA125 in predicting epithelial ovarian cancer: a meta-analysis

Authors:

Fake Li (lfk2346@yahoo.cn)
Ruxiu Tie (tieruxiu112@sina.com)
Kai Chang (changkai0203@163.com)
Feng Wang (wangfnew@gmail.com)
Shaoli Deng (dengsl072@yahoo.com.cn)
Weiping Lu (ronnylu@126.com)
Lili Yu (liliyuyu@sina.com)
Ming Chen (chenming1971@yahoo.com)

Version: 2 Date: 2 May 2012

Author's response to reviews: see over
Dear Prof. Sophie F. Derchain,

We would like to submit our revised attached manuscript to BMC Cancer. Manuscript number: 9036077596580943. Version 1: “Does Risk for Ovarian Malignancy Algorithm excel human epididymis protein 4 and CA125 in predicting epithelial ovarian cancer: a meta-analysis”.

We thank the reviewers for their valuable and constructive comments. We have carefully considered them and our replies are attached. Additionally, Lili Yu is currently receiving a grant (NO. 81070505) from National Natural Science Foundation of China. For other authors none are declared. All authors have read and approved the final revised version. We hope the manuscript can be accepted for publication in BMC Cancer.

Sincerely yours,

Ming Chen M.D.   Corresponding author

Department of Clinical Laboratory, Institute of Surgery Research, Daping Hospital, Third Military Medical University.

10 Changjiangzhilu Road, Yuzhong District, Chongqing 400042, China.

Email: chenming1971@yahoo.com

Tel: (+86) 23-68757601

Fax: (+86) 23-68716530
A. RESPONSES TO COMMENTS OF EDITOR

Question 1). Structured Abstract - Please format your abstract according to the guidelines for authors <http://www.biomedcentral.com/info/ifora/abstracts>.

Answer: Thanks for your comments. I have added the Methods part to make a Structured Abstract, according to the guidelines for authors.

Revised manuscript

“Methods: Remote databases (MEDLINE/PUBMED, EMBASE, Web of Science, Google Scholar, the Cochrane Library and ClinicalTrial.gov) and full texts bibliography were searched for relevant abstracts. All studies included were closely assessed with the QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies-2). EOC predictive value of ROMA was systematically evaluated, and comparison among the predictive performances of ROMA, HE4 and CA125 were conducted within the same population. Sensitivity, specificity, DOR (diagnostic odds ratio), LR± (positive and negative likelihood ratio) and AUC (area under receiver operating characteristic-curve) were summarized with a bivariate model. Subgroup analysis and sensitivity analysis were used to explore the heterogeneity.”

B. RESPONSES TO COMMENTS OF REVIEWER1

(1682476435693354_comment)

Major Compulsory Revisions

Question 1). The authors should divide the manuscript into 2 major parts:

a. Performance of CA125/HE4/ROMA in Ovarian Cancer (OC)
b. Performance of CA125/HE4/ROMA in Epithelial Ovarian Cancer (EOC)
The performance of CA125/HE4/ROMA in subsets of patients should be left out of the manuscript since this is not a relevant question. In fact, at the moment the manuscript is confusing and the methods and results section lack transparency. When reading the manuscript it is never clear what kind of analysis is performed. All figures and tables should also be divided in 2 parts or, even better, use different tables for the different outcomes and split all tables/figures.

**Answer:** We appreciate your constructive advices. To make our manuscript more understandable, we have adjusted all essential parts of the whole manuscript. For example, a concise *Methods* part has been added in the *Abstract*, and a *Methods section in the text* (page 5) has been placed after the *Backgrounds* section. All figures and tables have been split into independent tables according to the outcomes.

We majorly talk about EOC in this meta-analysis. In some studies included in this meta-analysis, OC is investigated as a whole, while EOC is not specified. So we set a subgroup analysis for OC prediction. And there are only two studies (Van Gorp et al, 2011 and Jacob et al, 2011) investigate HE4, CA125 and ROMA simultaneously for OC prediction. This is not suitable for meta-analysis. Thus we haven’t divided the manuscript into two major parts. However, we have added the performance comparison of HE4 and CA125 for OC prediction in the revised version.

**Revised manuscript**

“Five studies [16, 33-36] with 883 patients compared the performance of HE4 and CA125 for OC prediction.”

**Question 2).** Table 1 only includes 5 studies. It is not clear to me why the other 6 studies are not mentioned. Please also put the studies in alphabetical or chronological order.

**Answer:** Thanks for your review. There ought to be 11 studies in table 1, however, for the reason of online format converting, cells containing the other 6 studies were out of the page margins and missing. We have adjusted the format of table 1 in a new
version. And all the studies have been list in chronological order.

**Question 3** It is not clear to me how the authors handle different cut-off values when they compare sensitivity and specificity between the different studies. E.g. for HE4 different cut-offs of 70 pM, 72 pM, 74.2 pM, etc. are used. The different cut-offs will influence the sensitivity and specificity to the extent that they are not comparable, and yet, the authors compare the different studies. The same can be said about ROMA (premenopausal 13.1 – 12.5 and postmenopausal 14.4 -27.7). I have severe statistical concerns about this. To my opinion, only AUC's can be compared because they do not depend on the different cut-offs.

**Answer:** I agree with you. The different cut-offs will influence the sensitivity and specificity, but do not influence the AUC. So it is unreasonable to compare the sensitivity and specificity among studies with different cut-offs. In this manuscript, the sensitivity and specificity of the including studies have been summerized via univariate meta-analysis model (with 3 or less studies) or bivariate model. The summary sensitivity, summary specificity and other estimates have been compared among tests of CA125, HE4 and ROMA. Before summerizing of the estimates, heterogeneity of sensitivity and specificity estimates and ‘threshold effect’ have been tested. If low heterogeneity ($I^2 < 25\%$) of sensitivity and specificity estimates exist, two univariate meta-analysis for sensitivity and specificity estimates are used. If moderate ($25\% \leq I^2 < 50\%$) to high ($I^2 \geq 50\%$) heterogeneity of sensitivity and specificity estimates exist, the ‘threshold effect’ is required to be tested. The term ‘threshold effect’ is for the situation when some or all of the variation between studies can be explained by differences in the threshold among the studies. If the threshold effect presents, a positive Spearman correlation coefficient $\rho$ appears, thus the bivariate model is used to summerize the estimates (specificity, sensitivity and others). However, the coefficient $\rho$ may be close to zero or even negative. If $\rho$ is zero, two univariate meta-analysis for sensitivity and specificity estimates are used the same way as when low heterogeneity of sensitivity and specificity estimates exist. If
ρ is negative, bivariate model and interval estimates of parameters are described, but without a SROC curve. We have add essential information of our statistical approach in the part of Data Analysis Plan, according to reference (Chappell FM, Raab GM, Wardlaw JM: When are summary ROC curves appropriate for diagnostic meta-analyses? Stat Med 2009, 28:2653-2668).

**Question 4)** Table 2: it is not clear to me why in most of the studies only one of the index tests is regarded as low concern for applicability and the others high concern for applicability, even when all tests are performed according to standard operating procedures. ROMA is actually being sold as a combination of 2 EIA’s!! Please look at: http://www.he4test.com/row/professionals/inserts.html So why is the CA125 EIA high concern and the HE4 EIA not?? EIAs are general practice. So the following question: Are there concerns that the index test, its conduct, or interpretation differ from the review question? Should be answered with low.

**Answer:** Thanks for your question. We have closely assessed the including studies according to the quality items in QUADAS-2. The results of the assessment have been shown in table 2. On the question “ROMA is actually being sold as a combination of 2 EIA’s”, we have a different idea. The Risk of Ovarian Malignancy Algorithm (ROMA™) has received clearance from the FDA of U.S. in September 2011. On the website of Fujirebio Diagnostics, Inc., the news on September 6, 2011 says that ROMA™ is a qualitative serum test combining the results of HE4 EIA, ARCHITECT CA 125 II™ and menopausal status into a numerical score. The HE4 array is EIA, but the CA125 array is a chemiluminescent Microparticle immunoassay (CMIA) on the ARCHITECT i System. In this manuscript, CMIA, EIA, RIA (radioimmunoassay), CLEIA (chemiluminescence enzyme immunoassay) and ECLIA (electrochemiluminescence immunoassay) have been employed to measure CA125 in studies included. CA125 arrays with EIA and RIA are assessed as high Concern Regarding Applicability.

**Question 5)** Results section: 2nd paragraph: only reference 15, 17, 18, 20, 22, 23 are
mentioned. Why are references 16, 19 and 21 not mentioned? Did they not contribute to the 2879 patients? Again, it is not clear whether you are discussing only EOC or also OC. Please clarify.

**Answer:** In the manuscript, we only show the studies we included for the relevant analysis, the studies not included and their reasons are shown as follows.

Of the 11 studies (ref. 13-23) included in the manuscript, 6 studies (ref. 15, 17-18, 20, 22-23) investigated the performance of ROMA for EOC prediction; among the other 5 studies (ref. 13-14, 16, 19, 21), 4 studies (ref. 13-14, 19, 21) did not investigate ROMA; 1 study (ref. 16) investigated the performance of ROMA for ovarian cancer (OC) prediction.

Of the 11 studies (ref. 13-23) included, 4 studies (ref. 13, 15, 18, 22) compared the performance of HE4 and CA125 for EOC prediction; among the other 7 studies (ref. 14, 16-17, 19-21, 23), 1 study (ref. 14) investigated premenopausal women; 4 studies (ref. 16, 19-21) compared the performance of HE4 and CA125 for OC prediction; 2 studies (ref. 17, 23) investigated the performance of ROMA for EOC prediction only.

Of the 11 studies (ref. 13-23) included, 3 studies (ref. 15, 18, 22) compared the performance among ROMA, HE4 and CA125 for EOC prediction; among the other 8 studies (ref. 13-14, 16-17, 19-21, 23), 1 study (ref. 16) compared the performance among ROMA, HE4 and CA125 for OC prediction; 4 studies (ref. 13-14, 19, 21) did not investigate ROMA; 2 studies (ref. 17, 23) investigated the performance of ROMA for EOC prediction only; and 1 study (ref. 20) did not investigate HE4 and CA125 for EOC prediction.

All the patients numbers are listed in table 1. The calculation for total numbers is (531+389+65+104+233+159+229+ [419-1] +118+160+472 = 2878). In the study (Bandiera et al, 2011), 419 patients were enrolled, however, 1 patient with unknown stage was not analyzed, so 418 patients were analyzed. Finally 2878 patients were included in this meta-analysis. So the total number of patients included in this meta-analysis is corrected as 2878.

All the 11 studies contribute to the 2878 patients. We majorly talk about EOC, while OC has been discussed for performance comparison of HE4 and CA125 for OC prediction.
prediction.

**Question 6)** Please always clarify in the results section which of the studies are being mentioned. For example: “Within 9 of 11 studies, the results interpretation of index tests and reference standard tests were blind with each other.” Which studies?? Another example: “When evaluating patients selection with QUADAS-2, 4 of the total 11 studies with consecutive enrollment were considered as low risk of bias, 2 studies with non-consecutive enrolling were regarded as high risk of bias and 5 studies was unclear.” Which studies?? This should be corrected in the whole manuscript.

**Answer:** Thanks for your careful review. The citations of studies have been corrected in the revised version.

**Revised manuscript**

“Quality of included studies was assessed by the QUADAS-2 tool (Fig. 2 & Table 2). Within 9 [13-14, 16-18, 34-37] of 11 studies, the results interpretation of index tests (HE4/CA125) were blind with reference standard test (ROMA). The other 2 studies [15, 33] were unclear. In 5 of the 11 studies [14, 16, 34-36] the results of index tests (HE4 and CA125) were interpreted without knowledge of each other. In the other 6 studies [13, 15, 17-18, 33, 37] the blindness was unclear. So when assessing the studies with the item “Could the conduct or interpretation of the index test have introduced bias?” in domain 2 of QUADAS-2, the results showed that 5 studies [14, 16, 34-36] were low risk of bias, 1 study [13] was high risk of bias and 5 studies [15, 17-18, 33, 37] were unclear their risk of bias. Four [16-18, 34] of the total 11 studies were considered as low risk of bias for the Patient Selection (Domain 1 of QUADAS-2) for their consecutive enrollment of patients; 2 studies [14-15] were
regarded as high risk of bias and in the other 5 studies [13, 33, 35-37] the risk was unclear.”

**Question 7)** Figure 3: please provide the reader also with summary forest plots, not only the forest plots of the individual studies

**Answer:** We accept your suggestion. All summary value of the individual studies have been added in the forest plots in the revised version.

**Question 8)** Figure 4: please provide information on the studies depicted in this summary ROC graph + provide whether OEC or OC was the outcome

**Answer:** We are pleased to accept your suggestion. Figure 4 depicts the performance of ROMA predicting EOC. So figure legend text of Figure 4 has added as “Results of bivariate analysis: estimates of each studies (the squares), the summary point (solid circle), 95% confidence region (the small ellipse), 95% prediction region (the big ellipse) and HSROC (solid line) were shown. Each study is represented by each square in the meta-analysis. The size of the square indicates the size of each study.” Figure title has been rewrited as “Figure 4 - Hierarchical summary receiver operating characteristic (HSROC) curves and results of bivariate analysis for ROMA to predict EOC”

**Question 9)** Figure 5: again: outcome EOC or OC?

**Answer:** Thanks for your suggestion. The outcome in Figure 5 is EOC. The title of Figure 5 has been rewrited as “Figure 5 - Influence analysis of individual studies for performance of ROMA to predict EOC”.

**Question 10)** The authors fail to mention a comparison between the summary AUC of CA125, HE4 and ROMA with corresponding 95%CI and whether the differences are statically different. The authors refer to Table 3, but this table is very confusing and they do not provide p-values or differences in AUc with the corresponding 95%CI.
This information is crucial for the discussion and conclusion whether the use of ROMA will increase the diagnostic accuracy

**Answer:** Thanks for your question. We fail to make the Table 3 informative in the manuscript. And in the revised version, the Table 3 has been divided into three individual tables (Table 3, 4 & 5 in the revision) to be clearer. In the manuscript, we compared the summary AUCs of CA125, HE4 and ROMA in three studies, each of which employed CA125, HE4 and ROMA in the same population. And we found the AUC values were similar among the three tests (Table 5 in the revision). Their p values of comparing every two tests were all greater than 0.05.

**Question 11)** Methods section: LMP tumors and borderline tumors are the same. LMP tumors should not be considered as early stage ovarian cancer. LMP tumours can present as advanced stage as well.

**Answer:** Thanks for your review. I agree with your idea. Some of the studies included in the meta-analysis contain LMP tumors while some studies use the term borderline tumors. So in this manuscript we list both the term. It’s not reasonable to regard LMP tumors as early stage ovarian cancer. This has been corrected in the revised version.

**Revised manuscript**

“Early stage were defined as FIGO stages I & II, while advanced stage were FIGO stages III & IV.”

“In some studies, patients with low malignant potential tumors (LMP) or borderline tumors (BL) were classified into EOC group. And these studies were specifically analyzed as subgroup EOC (LMP/BL).”

**Minor Essential Revisions**

**Question 1)** Background: please clearly mention that CA125 is FDA approved, not for the diagnosis of ovarian cancer, but for monitoring disease response.

**Answer:** I entirely agree with you. The sentence involved in the Background part “Cancer antigen 125 (CA125) was the only FDA-approved biomarker for ovarian
cancer before the year 2008.” has been rewriter as “Cancer antigen 125 (CA125) was the only FDA-approved biomarker for monitoring ovarian cancer response to treatment before the year 2008.”

**Question 2)** Background: CA125 is also less expressed in mucinous tumours. Please correct the third sentence of the second paragraph.

**Answer:** Thanks for your review. We have corrected the mistakes in the second paragraph of Background and other places all over this manuscript.

**Revised manuscript**

“Unlike CA125, HE4 doesn’t overexpress in endometriosis and other benign gynaecological diseases.”

### C. RESPONSES TO COMMENTS OF REVIEWER2

(7336545196937456_comment)

**Major Compulsory Revisions**

**Question 1).** The study objectives and the link between these objectives and the study methods should be better explained. The last sentence of the background indicates that the aim is twofold:

I. assessing the accuracy of ROMA

II. comparing the accuracy of ROMA, HE4 and CA125 against each other.

**1a.** For the second question, the authors limit their study base to studies comparing either HE4 against CA125 or ROMA against both HE4 and CA125. Although this makes sense, I would like to know why the authors did not include studies evaluating only CA125 or only HE4.

**1b.** The authors do not explain in their methods section how they will analyze the comparative question. Which studies will feed into that (only comparative studies, or also the only-ROMA studies)? Did they check whether all samples were tested for
HE4 and for CA125 and whether all samples were included in the analyses of ROMA? To give an example, a study that included 100 patients of which 75 were sampled for CA125 and 85 for HE4, but only 60 patients delivered a sample for both tests, then only this subsample of 60 can be used for ROMA-analyses. I suspect that this is not an issue in this review (as it is easy to do two tests on one blood sample), but I can also imagine that sometimes not enough blood could be drawn for two test, or that CA125 was already included in the clinical routine (because approved by the FDA) and thus used more often.

1c. The authors used both QUADAS and QUADAS-2, but they did not include questions to assess the validity of this comparative question. Examples are: were the results for CA125 masked when HE4-result was read? Did all patients undergo both HE4, CA125 and ROMA? Were the patients in these studies the same as the patients in the studies only addressing HE4/CA125/ROMA?

**Answer:** Thanks for your constructive suggestions. We accept your suggestion and modify the Methods part to strengthen the linkage between the objectives and the study methods.

**Answer to 1a:** We can understand your concern. For diagnosis meta-analysis, the heterogeneity between the including studies inevitably exists. And the heterogeneity of the population between studies usually plays an important role in the overall heterogeneity. The smaller heterogeneity between the including studies ensures the more credible results of the meta-analysis. In this manuscript, only studies conducting both the two tests (HE4 and CA125) or all three tests (HE4, CA125 and ROMA) in a same population have been included in comparisons among tests. This makes sure that the comparison takes place between studies under the same or similar population background, thus reduces the heterogeneity between studies [42]. Take comparison between HE4 and CA125 for example, every participant in each of the including studies has been tested HE4 and CA125 simultaneously or tandem. In each study, HE4 or CA125 test has the same population background, therefore, for a meta-analysis of these studies, the HE4 test or CA125 test has same or similar
population background at least. So it is with comparison among HE4, CA125 and ROMA. In the routine diagnosis meta-analysis paper, each index test is usually literature searched, reviewed and summarized alone. Then the summary estimates (for instance, sensitivity) are compared between index tests, which often take place in different studies enrolling different participants. Thus the heterogeneity between studies would be prominent. So in this manuscript, studies evaluating only CA125 or only HE4 are not included.


Answers to 1b: (sentence by sentence)

“The authors do not explain in their methods section how they will analyze the comparative question.”

Answer to this sentence: In the Methods section we add the description of the approach to analyze the comparative question, as “Summary estimates and 95% CIs (confidence intervals) for sensitivity, specificity, DOR, LR± and AUC were calculated (STATA version 10.0 [31-32]). HSROC (Hierarchical summary receiver operating characteristic curves) plots were shown when appropriate. Comparisons between estimates of different tests were performed with z-test.”

“Which studies will feed into that (only comparative studies, or also the only-ROMA studies)?”

Answer to this sentence: In the manuscript, studies that investigated both serum HE4 and CA125 as diagnostic tests or calculated the ROMA algorithm were included. Some studies investigated both serum HE4 and CA125 to explore and compare their diagnostic roles, although didn’t calculate the ROMA algorithm, were included. And some studies investigated both serum HE4 and CA125 just to calculate the ROMA algorithm, but did not explore the diagnostic performance of HE4 or CA125, were also included.
“Did they check whether all samples were tested for HE4 and for CA125 and whether all samples were included in the analyses of ROMA?”

Answer to this sentence: We appreciate your careful consideration. We have checked the numbers of participants that were tested for HE4, CA125 and ROMA, they are accordant.

**Answers to 1c:** We appreciate your suggestions. The concerns of assessing the validity of this comparative question are essential. We have added these suggestions into quality items of QUADAS-2.

**Revised manuscript**

“In the items of QUADAS-2, the blindness of index tests and reference test has been list, but not the blindness between index tests. So one item that focus on validity of this comparative question has been added in Risk of Bias part of Domain 2 (Index Test) in QUADAS-2 [22] as follows. “Were the results of index tests interpreted without knowledge of each other?” The answers (Yes, No or Unclear) of this question were considered to help assessing the Risk of Bias of including studies.”

“According to the suggestion in Concerns Regarding Applicability part of Domain 2 (Index Test) in QUADAS-2 [22], variations in test technology, executing, or interpretation might affect estimates of the diagnostic accuracy of a test. If index test methods varied from those specified in the review question, concerns about applicability might exist.”

**Question 2).** The questions above reveal another problem with this review: the clinical context. The authors state that CA125 and ROMA are approved by the FDA, but they don’t explain whether these tests are used in practice and at which stage in the diagnostic process they are used. Neither do they explain this for HE4. What will these tests be used for and in what setting? Will they be used by the general practitioner to refer patients to the hospital (when test positive) or to send them home (when test negative)? Or is the test used to decide who goes for further diagnosis by,
for example, CT scanning? Or will these test be used to confirm the diagnosis of EOC?

**Answer:** I accept your suggestions. We have added the descriptions about the usable condition of CA125, HE4 and ROMA as follows in the Background section.

**Revised manuscript**

“CA125 is indicated for use as an aid in the detection of residual ovarian carcinoma in patients who have undergone first-line therapy and would be considered for diagnostic second-look procedures.”

“And HE4, as an aid in monitoring recurrence or progressive disease in patients with epithelial ovarian cancer, has been the first biomarker for EOC after CA125 to be approved by the U.S. Food and Drug Administration (FDA) at the year of 2008.”

“Moore and colleagues [17] have explored a multianalytes assay named the Risk of Ovarian Malignancy Algorithm (ROMA™), which combines the results of HE4 EIA (enzyme immunoassay), ARCHITECT CA 125 II™ and menopausal status into a numerical score to predict malignancy when an ovarian mass was found clinically. ”

**Question 3**. A comparative question also almost asks for a null hypothesis: what do the authors expect to find? That ROMA is more sensitive than both CA125 and HE4 alone? And when would the authors decide that ROMA is ‘better’ than either CA125 or HE4?

**Answer:** Thanks for your questions. In this manuscript, we intend to find whether ROMA is better than HE4 and CA125 in predicting EOC. Thus, the diagnostic accuracy of ROMA is explored, and the performance of ROMA, HE4 and CA125 are compared. In our manuscript, we pool studies investigating HE4, CA125 and ROMA simultaneously. We find that ROMA is more sensitive than HE4, but similar to CA125; ROMA is less specific than HE4 but more specific than CA125. In the design of diagnostic meta-analysis, one test has a higher AUC is better than the other. However, in the results of pooling 3 studies, no differences are found between the AUCs of HE4, CA125 and ROMA. The ROMA is judged better than CA125, due to their specificity difference (ROMA: 0.84, 95%CI 0.79-0.88 vs CA125: 0.78, 95%CI 0.73-0.83).
**Question 4.** Did the authors use the studies that compared HE4 with CA125 to calculate ROMA from these results and include them in the ROMA-analysis?

**Answer:** We only include the studies with calculated ROMA score in the study pooling of the ROMA-analysis. Because the original data of the HE4 value and CA125 value of each participant, we couldn’t calculate the ROMA score.

**Question 5.** The authors state that they found 11 articles meeting the inclusion criteria. But from the text of the search results section, it appears that the three studies that compared the performance of ROMA, CA125 and HE4, were among the 6 only evaluating ROMA. Also, of the four studies comparing HE4 and CA125, three were already mentioned under evaluation of ROMA alone. So I come to 7 articles in stead of 11. Also, the numbers of patients seem to be not correct (some patients counted twice?).

Under the methods of index tests section, the numbers are different. Eight studies used EIA to measure HE4. The references 16 and 19 and 21 were not mentioned under the previous section. Three other studies used CMIA, but reference 14 was not mentioned under the previous section. This implies that the authors did search for studies only including HE4. The same seems to be true for CA125. This is all very confusing!

This is all very confusing and a clearer explanation of what analyses were included in which studies/articles is really needed.

**Answer:** Thanks for your careful review. In the manuscript, indeed 11 articles meet the inclusion criteria. The 7 studies (13,15,17,18,20,22,23) (Bandiera et al, 2011; Jacob et al, 2011; Kim et al, 2011; Montagnana et al, 2011; Moore et al, 2008; Moore et al, 2009; Moore et al, 2011.) you count are on the topic of EOC, while the three studies (16, 19, 21) (Abdel-Azeez et al, 2010; Chang et al, 2011; Van Gorp et al, 2011.) are on the topic of Ovarian Cancer (OC) , and the study (14) (Holcomb et al, 2011) is also on the topic of OC. All the four studies (14, 16, 19, 21) (Abdel-Azeez et al, 2010; Chang et al, 2011; Holcomb et al, 2011; Van Gorp et al, 2011.) are searched and
included via the same criteria as the other 7 studies. All the patients numbers are listed in table 1 (531+389+65+104+233+159+229+[419-1]+118+160+472=2878), none patients have been counted twice. In the study (Bandiera et al, 2011), 419 patients were enrolled, however, 1 patient with unknown stage was not analyzed, so 418 patients were analyzed. Finally 2878 patients were included in this meta-analysis.

We have added the descriptions about OC investigation as follows in the Methods section to make the included studies clearer.

**Revised manuscript**


“The study of Holcomb and colleagues[14] had the lowest prevalence (7.86%) with only results of premenopausal women. Thus it was included in the subgroup of premenopausal women when comparing the performance of HE4 and CA125 for EOC monitoring.”

**Question 6**. The authors state that “in all studies, the spectrum of patients was considered representative”, but they do not explain what this means. Are these women coming to primary care, secondary care, etc?

**Answer**: Thanks for your suggestion. We have modified this sentence “in all studies, the spectrum of patients was considered representative” in the revised version to make it more understandable. The sentence are modified as “In all studies, the spectrum of patients was considered representative. All enrolled participants present pelvis mass of suspected ovarian origin, have never received any treatment before and plan to have a surgical intervention.”

**Question 7**. I don’t understand the explanation about the concerns of applicability for domain 2 of QUADAS-2. Please add a few sentences of what this really means.

**Answer**: We have added more description about QUADAS-2 including the
explanation about the concerns of applicability (domain 2 of QUADAS-2). The explanation for concerns of applicability (domain 2 of QUADAS-2) are “As Domain 2 (Index Test) of QUADAS-2 suggested, variations in test technology, executing, or interpretation may affect estimates of the diagnostic accuracy of a test. If index test methods vary from those specified in the review question, concerns about applicability may exist. For HE4 arrays, the chemilumenesence immunoassays (CLIA) are more sensitive than EIA (specified assay for HE4), thus bias may be introduced into pooling of studies with CLIA. And similarly, for CA125, EIA and RIA assays are less sensitive and steady than CLIA (specified assay for CA125), so studies using either EIA or RIA will be considered as High Concern Regarding Applicability. The ROMA test employed both the results of CA125 and HE4 within the same study. So ROMA was considered as High Concern Regarding Applicability when either HE4 or CA125 test was evaluated as High Concern Regarding Applicability.”

**Question 8).** The results under the ROMA diagnostic value section seem to be analysed correctly. But I do miss an explanation of which studies feed into which analyses (although it can be extracted from the forest plots). I also think that there are too many forest plots to be informative. I would rather restrict to only one forest plot and mention the results for the subgroups in the text only.

**Answer:** Thanks for your suggestion. I will add the explanation that which studies feed into which studies in the Results section of the revised version. I accept your constructive suggestion to display only one forest plot and mention the results for the subgroups in the tables.

**Question 9).** The results under the two performance comparison sections also seem to be correct. But I don’t understand how it is possible that only four studies compared HE4 to CA125, while all 11 studies evaluating HE4, also evaluated CA125 (see under methods of index test). Again, please explain better which studies included which analyses and how these analyses were used in the meta-analyses and why.

**Answer:** In the manuscript, we only show the studies we included for the relevant
analysis, the studies not included and their reasons are shown as follows.


2011; Montagnana et al, 2011; ) compared the the performance among ROMA, HE4 and CA125 for EOC prediction; among the other 8 studies (ref. 13-14, 16-17, 19-21, 23) (Abdel-Azeez et al, 2010; Chang et al, 2011; Holcomb et al, 2011; Kim et al, 2011; Moore et al, 2008; Moore et al, 2009; Moore et al, 2011; Van Gorp et al, 2011), 1 study (ref. 16) (Van Gorp et al, 2011) compared the the performance among ROMA, HE4 and CA125 for OC prediction; 4 studies (ref. 13-14, 19, 21) (Abdel-Azeez et al, 2010; Chang et al, 2011; Holcomb et al, 2011; Moore et al, 2008; ) did not investigate ROMA; 2 studies (ref. 17, 23) (Moore et al, 2009; Moore et al, 2011) investigated the performance of ROMA for EOC prediction only; and 1 study (ref. 20) (Kim et al, 2011) did not investigate HE4 and CA125 for EOC prediction.

**Question 10.** I miss a sROC plot in which all tests are included. Or at least a separate one for CA and HE.

**Answer:** Thanks for your suggestion. We have added two HSROC plots (Fig. 8 & 11) including HE4 and CA125 for their EOC and OC prediction and a sROC plot (Fig. 15) including all tests (ROMA, HE4 and CA125) for their EOC prediction.

**Question 11.** Would it be possible to formally test for a difference in the accuracy of ROMA versus HE4 versus CA125? Under the Conclusions section, the authors state that ROMA and HE4 can replace CA125. But the confidence intervals of ROMA and CA125 show very much overlap and I can’t find the sensitivity of CA125 (so can’t judge how different this is from the other two).

**Answer:** In this manuscript, three studies investigate ROMA, HE4 and CA125 simultaneously. Based on the comparison of the three studies (Bandiera et al, 2011; Jacob et al, 2011; Montagnana et al, 2011), in which accuracy of ROMA, HE4 and CA125 are compared two by two. And the significant difference (p< 0.05) is labeled with symbol characters, such as “★● 〓 ★● 〓 ★● 〓 ★● 〓 # # ## * * **”. In table 5, no significant difference exists in the sensitivity of ROMA (0.86, 95%CI 0.81-0.91) and CA125 (0.84, 95%CI 0.78-0.89). And the AUCs (area under SROC curve) of ROMA (0.92, 95%CI 0.86-0.97), HE4 (0.95, 95%CI 0.92- 0.98) and CA125 (0.88, 95%CI 0.81- 0.96) are similar.
But both the specificity of HE (0.94, 95% CI 0.90-0.96) and ROMA (0.84, 95% CI 0.79-0.88) are higher than CA125 (0.78, 95% CI 0.73-0.83) (p<0.05). So ROMA and HE4 are better (specificity) than CA125 in predicting EOC for participants with pelvic mass. However, based on the results of comparison of HE4 and CA125 (4 studies), we find that the AUCs of HE4 are lower than CA125 for EOC (HE4 0.82, 95% CI 0.78-0.85 vs CA125 0.88, 95% CI 0.85-0.91) or OC (HE4 0.79, 95% CI 0.76-0.83 vs CA125 0.89, 95% CI 0.85-0.91) prediction. So more studies with high quality are needed to included in the meta-analysis in the future. To make the conclusion more precise, we have rewrited the Conclusions section as “ROMA can help distinguish EOC from benign pelvic mass. ROMA is less specific but more sensitive than HE4. Both ROMA and HE4 are more specific than CA125 for EOC prediction. CA125 has better diagnosis accuracy than HE4 for EOC and OC prediction. ROMA is promising predictor to replace CA125, but its utilization requires further exploration.” And Conclusions in the Abstract: “ROMA is helpful for distinguishing epithelial ovarian cancer from benign pelvic mass. HE4 is not better than CA125 either for EOC of OC prediction. ROMA is promising predictors of epithelial ovarian cancer to replace CA125, but its utilization requires further exploration.”

**Question 12.** Table 3 is unreadable and impossible to understand. This may be due to the format. Also, what do the numbers under the comparative section of table 3 mean? If it says 0.86 under sensitivity for EOC-ROMA in the ROMA:HE:CA section, how does this differ from the 0.90 for the only-ROMA sensitivity (should perhaps be addressed in the discussion)?

**Answer:** Thanks for your suggestion. We fail to make the Table 3 informative in the manuscript. In order to be clearer, the Table 3 has been divided into three individual tables (Table 3, 4 & 5) in the revised version.

The difference between 0.86 (the sensitivity for EOC-ROMA in ROMA:HE4:CA125 section, n=3) and 0.90 (the sensitivity for the only-ROMA, n=6) majorly exists in the study numbers. The former includes 3 studies while the later
includes 6. So they are with the other estimates (specificity, DOR, LR± and AUC). Although they are all for ROMA, they are in different settings. When the performance of ROMA for EOC prediction is talk about, the estimates of ROMA only are used. Meanwhile, when ROMA is compared with HE4 or CA125, only estimates gained from the ROMA vs HE4 vs CA125 are used. In the revised version, each estimates with their values were writed clearly.

**Question 13**. Please address in the discussion what these results mean and for which patients. What does it mean for example, an AUC of 0.93 for ROMA? What do all the I-squares in table 3 mean; what is the meaning of some being 0.0%, while others are 88%? What is the consequence of ROMA being better capable of detecting advanced stage EOC than detecting early stage EOC (while the specificity remains the same)? What do the numbers for the comparative analyses mean (see also table 3)?

**Answer**: Thanks for your suggestions. We have described the estimates of every test and test comparison in the Results section. In the Discussions section, a summary of the results have been shown.

**Revised manuscript**

“Our results found that, first, ROMA could help distinguish EOC from benign pelvic mass with a high diagnostic accuracy (AUC: 0.93). The ROMA has high sensitivity to predict advanced stage EOC than early stage EOC and in postmenopausal women than in premenopausal women. Second, although HE4 has higher specificity than CA125 for EOC monitoring, CA125 has better diagnosis accuracy (higher AUC) than HE4 for EOC or OC prediction. This is based on the results of 4 studies that compare HE4 and CA125 within the same population. Third, based on the results of comparison of HE4, CA125 and ROMA in the same population, the overall performance (AUC) of the three tests for EOC prediction are similar. ROMA is less specific but more sensitive than HE4, while both ROMA and HE4 are more specific than CA125 for EOC monitoring.”

The explanation of I-squares in table 3 has been added in the Methods section as follows.
Revised manuscript

“Heterogeneity of studies included were shown with forest graphs and explored with $I^2$ estimates [27]. The main advantage of $I^2$ was that it was inherently independent with the number of the studies in the meta-analysis. $I^2$ estimates below 25% were regarded as low risk of heterogeneity, between 25% and 50% as moderate heterogeneity, and 50% or higher as high heterogeneity.”

Although the specificity of ROMA in EOC-early stage and EOC-advanced stage remains the same, the sensitivity and DOR of ROMA in EOC-advanced stage are higher than in EOC-early stage. In EOC-advanced stage: sensitivity 0.98 (0.94- 1.00), DOR 149.08 (47.80- 464.95). In EOC-early stage: sensitivity 0.81 (0.71- 0.89), DOR 17.18 (9.08- 32.50).

I don’t think I really catch the meaning of the question “What do the numbers for the comparative analyses mean (see also table 3)” My answer is that the “n” in tables means numbers of the studies that are suitable the corresponding clinical setting.

Question 14). Please add a short description of the methods to the abstract.
Answer: A short description of the Methods part has been added to the Abstract section.

Revised manuscript

“Methods: Remote databases (MEDLINE/PUBMED, EMBASE, Web of Science, Google Scholar, the Cochrane Library and ClinicalTrial.gov) and full texts bibliography were searched for relevant abstracts. All studies included were closely assessed with the QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies-2). EOC predictive value of ROMA was systematically evaluated, and comparison among the predictive performances of ROMA, HE4 and CA125 were conducted within the same population. Sensitivity, specificity, DOR (diagnostic odds ratio), LR± (positive and negative likelihood ratio) and AUC (area under receiver operating characteristic-curve) were summarized with a bivariate model. Subgroup
analysis and sensitivity analysis were used to explore the heterogeneity.”

**Question 15.** The authors state that they used the bivariate model to analyze sensitivity and specificity. The studies report however a wide variety of cut-off values (for HE and ROMA). What does a summary sensitivity and specificity mean then? To which cut-off value does this relate? Would it matter if someone uses one cut-off value or another? In general, when studies report varying cut-off values, the HSROC model is recommended. State reports the results for this model as well (although the results may be more difficult to interpret). Why didn't the authors use the results from this model?

**Answer:** Thanks for your questions. The different cut-offs will influence the sensitivity and specificity, but do not influence the AUC. So it is unreasonable to compare the sensitivity and specificity among studies of different cut-offs.

Bivariate methods *(Reitsma JB, Glas AS, Rutjes AWS, Scholten RJPM, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. Journal of Clinical Epidemiology 2005; 58:982-990.)* and the hierarchical SROC methods *(Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Statistics in Medicine 2001; 20:2865-2884.)* are two random-effects methods that have been developed to overcome the limitations of the Moses-Littenberg method and permit studies to vary in both threshold and accuracy.

In this manuscript, the sensitivity and specificity of the including studies have been summerized via either univariate meta-analysis model or bivariate model. The summary sensitivity, summary specificity and other estimates have been compared among tests of CA125, HE4 and ROMA. Before summerizing of the estimates, heterogeneity of sensitivity and specificity estimates and ‘threshold effect’ have been tested. If low heterogeneity of sensitivity and specificity estimates exist, two univariate meta-analysis for sensitivity and specificity estimates are used. If low heterogeneity of sensitivity and specificity estimates exist, the ‘threshold effect’ is required to be tested. The term ‘threshold effect’ is for the situation when some or all
of the variation between studies can be explained by differences in the threshold among the studies. If the threshold effect present, a positive correlation coefficient $\rho$ appears, thus the bivariate model is used to summerize the estimates (specificity, sensitivity and others). However, the estimate of $\rho$ may be close to zero or even negative. If $\rho$ is zero, two univariate meta-analysis for sensitivity and specificity estimates are used the same way as when low heterogeneity of sensitivity and specificity estimates exist. If $\rho$ is negative, we describe bivariate model and interval estimates of parameters, but without SROC curve. We have add essential information of our statistical approach in the part of Data Analysis Plan, according to reference (Chappell FM, Raab GM, Wardlaw JM: When are summary ROC curves appropriate for diagnostic meta-analyses? Stat Med 2009, 28:2653-2668). Both bivariate model and HSROC model require 4 studies at least, so the comparison among HE4, CA125 and ROMA including only 3 studies is investigated via univariate meta-analysis model with SROC curves. HSROC plot of the performance of ROMA for EOC prediction has been shown in the manuscript. We have also added HSROC plots including both HE4 and CA125 for EOC or OC prediction and a SROC plot including all tests (ROMA, HE4 and CA125) for EOC prediction in the revised version.

**Minor Essential Revisions**

**Question 1)** The methods section comes after the results section, which is very confusing and unhelpful.

**Answer:** Thanks for your review. The Methods section has been placed after the Backgrounds section in the revised version.

**Question 2)** Where does OC stand for and what is the meaning of the subgroup analyses for OC?

**Answer:** OC stands for ovarian cancer, which contains epithelial ovarian cancer (EOC). We majorly talk about EOC in this meta-analysis. In some studies included in this meta-analysis, OC is investigated as a whole, while EOC is not specified. So we set a subgroup analysis for OC prediction. And OC has also been discussed for
performance comparison of HE4 and CA125 for OC prediction in the revised version.

D. RESPONSES TO COMMENTS OF REVIEWER3

(6688864506940220_comment)

Minor Essential Revisions

Question 1) p. 5 – sentence “Applicability when either HE4 or CA125 test was evaluated as.” Is not finished

Answer: Thanks for your attentive review. I have rewrite this sentence to make it complete. The revised sentence has been marked red in the revised manuscript. “The ROMA tests were considered as high Concern Regarding Applicability when either HE4 or CA125 test was evaluated as high Concern Regarding Applicability (Fig. 2).”

Question 2) Next sentence – typo

Answer: I agree with your suggestion. This sentence is not necessary here. This sentence has been placed in the legend text of figure 2b in the revised manuscript. The changes have also been marked red. “Fig 2b. Proportion of studies with low, high, or unclear Concerns Regarding Applicability. Three horizontal bars represented index tests HE4, CA125 and ROMA, respectively.”