Serum human epididymis protein 4 vs carbohydrate antigen 125 for ovarian cancer diagnosis: a systematic review

Simona Ferraro,[1] Federica Braga,[1] Monica Lanzoni,[2,3] Patrizia Boracchi,[2,3] Elia Mario Biganzoli,[2,3] Mauro Panteghini[1]

▶ Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/jclinpath-2012-201031).

[1]Laboratorio Analisi Chimico-Cliniche, Azienda Ospedaliera 'Luigi Sacco', and Cattedra di Biochimica Clinica e Biologia Molecolare Clinica, Università degli Studi, Milano, Italy
[2]Sezione di Statistica Medica e Biometria, Università degli Studi, Milano, Italy
[3]Fondazione IRCCS Istituto Nazionale dei Tumori, Milano, Italy

Correspondence to
Dr Simona Ferraro, Laboratorio Analisi Chimico-Cliniche, Ospedale 'Luigi Sacco', Via G.B. Grassi 74, Milano 20157, Italy; ferraro.simona@hsacco.it

Received 27 June 2012
Revised 7 December 2012
Accepted 13 December 2012
Published Online First 20 February 2013

To cite: Ferraro S, Braga F, Lanzoni M, *et al*. *J Clin Pathol* 2013;**66**:273–281.

ABSTRACT

Background Human epididymis protein 4 (HE4) measurements in serum have been proposed for improving the specificity of laboratory identification of ovarian cancer (OC).

Objective To critically revise the available literature on the comparison between the diagnostic accuracy of HE4 and carbohydrate antigen 125 (CA-125) to confirm the additional clinical value of HE4.

Methods A literature search was undertaken on electronic databases and references from retrieved articles; articles were analysed according to predefined criteria. Meta-analyses for HE4 and CA-125 biomarkers with OR, diagnostic sensitivity, specificity, positive (LR+) and negative (LR–) likelihood ratios as effect sizes were performed.

Results 16 articles were originally included in meta-analyses, but two for HE4 and one for CA-125 were eliminated as outliers. Furthermore, for HE4 a publication bias was detected. ORs for both HE4 (37.2, 95% CI 19.0 to 72.7, adjusted for publication bias) and CA-125 (15.4, 95% CI 10.4 to 22.8) were significant, although in a heterogeneous set of studies ($p<0.0001$). By combining sensitivity and specificity, the overall LR+ and LR– were 13.0 (95% CI 8.2 to 20.7) and 0.23 (95% CI 0.19 to 0.28) for HE4 and 4.2 (95% CI 3.1 to 5.6) and 0.27 (95% CI 0.23 to 0.31) for CA-125, respectively.

Conclusions HE4 measurement seems to be superior to CA-125 in terms of diagnostic performance for identification of OC in women with suspected gynaecological disease. Due to the high prevalence of OC in post-menopausal women and the need for data focused on early tumour stages, more studies tailored on these specific subsets are needed.

INTRODUCTION

Ovarian cancer (OC) is the sixth most common gynaecological malignancy characterised by an incidence rate that increases with age and in post-menopausal status. The crude incidence rate changes from 4.7 per 100 000 in women <50 years of age to 29.6 per 100 000 in the age group of 50–64 years.[1] OC is currently the first cause of death in gynaecological malignancies; ~75% of patients are diagnosed at an advanced stage, since OC is generally asymptomatic in the early stages and no effective screening approach is available.[2] The net discrepancy between survival rates in early and advanced stages (80–90% vs 15–20%) has reinforced the need for biomarkers with higher diagnostic accuracy to set up screening programmes and/or to early distinguish malignancy from benign pelvic mass.[2]

Carbohydrate antigen 125 (CA-125) is the established biomarker for detecting OC recurrence and monitoring therapeutic response. In addition, recent guidelines recommend its measurement in the primary care setting in women with suggestive symptoms or at high risk for OC, in combination with pelvic ultrasound,[3 4] even though some authors have discouraged this application because of the low sensitivity of the test, which is even worse in early stage tumours (~50%).[5] It is noteworthy that CA-125 is consistently expressed in serous and endometrioid OC, whereas tumours detectable at early stages have a higher prevalence of non-serous carcinomas.[6] Overall, CA-125 effectiveness in the identification of the malignancy is threatened by its low diagnostic specificity. In fact, this glycoprotein is widely distributed on the surface of cells of mesothelial origin in various benign and malignant conditions other than OC.[7]

Among a wide spectrum of biomarkers recently proposed to aid in the diagnosis of women with suspected OC,[8] human epididymis protein 4 (HE4) is undoubtedly the most promising. Its measurement was from the beginning proposed to improve the diagnostic specificity of CA-125, just maintaining a similar sensitivity.[9] HE4 has homology with some secreted serine protease inhibitors and was reported to be amplified in some CA-125-deficient OCs, whereas its expression is lower in normal ovarian tissue, ovarian benign disease and low-malignant potential tumours.[10] After preliminary studies confirming genomic and immunohistochemical findings on HE4,[8] a large body of literature has been recently produced. Despite the low number of initially available studies, recent guidelines resorting to a meta-analytic approach have suggested HE4 to be used as an aid in OC diagnosis.[3] In addition, a systematic review (SR) has been recently published reporting better diagnostic performance in terms of sensitivity, specificity and likelihood ratios (LR) for HE4 than for CA-125.[11] However, the type of included studies, the applied selection criteria and the statistical approach used to synthesise the evidence could be criticised. Exploiting the more recent increase of studies on the comparison of HE4 and CA-125 diagnostic performances for OC, we designed an SR to critically revise available literature overcoming the above-reported threats of Yu's SR. In particular, we sought to provide a synthesis of the available evidence on the diagnostic accuracy of the tests by considering only those

studies evaluating both markers on the same case series. Methodologically, a stepwise selection of the studies and a further application of proper summary receiving operating curves (SROC) analysis was used to strengthen the evidence.

METHODS

Literature search strategy for identification of studies

The peer-reviewed literature published up to January 2012 was searched using the Medline (since 1966) and Embase (since 1993) databases, with Mesh terms (Human Epididymis 4 or HE4 and Ovarian), and with limits 'Title/Abstract, Human Subjects, English'. In addition, the reference lists of retrieved articles and of a previously published meta-analysis were screened to identify further studies.[11] The final aim of the search was to identify those original articles in which serum/plasma HE4 and CA-125 measurements were investigated and compared for OC diagnosis in order to provide a synthesis of the scientific evidence by the meta-analysis process.

Article evaluation and data extraction

First, two reviewers (SF and FB) evaluated the title and abstract of all preliminary identified records to assess whether the paper was relevant to the aim of the study. Then, by evaluating the complete manuscript, it was determined whether the preliminary selected papers met the following main criteria:

1. The primary or secondary aim of the study was at least the report of HE4 and CA-125 mean concentrations or sensitivity and specificity versus the 'gold standard' method for OC diagnosis—that is, laparoscopy with histological evaluation of biopsy material.
2. Diagnostic parameters were estimated according to a decisional threshold level and not to a fixed specificity or sensitivity.
3. The presentation of quantitative data allowed at least calculation of the OR.
4. The investigated population was represented by women with a gynaecological disease suspected as being OC, which is the intended spectrum of patients to be investigated by circulating biomarker detection.
5. HE4 concentrations were not included in diagnostic algorithms used to classify patients (incorporation bias).

Papers were excluded in the following instances:

1. Duplicative results from the same authors' group were being reported.
2. Serum/plasma HE4 concentrations were measured to assess OC recurrence, to monitor disease progression or the effect of therapy.
3. They were case reports.

The quality of the selected studies was judged according to the QUADAS II [12] criteria.

This four-phase tool was built first to identify possible sources of bias concerning patient selection, index test, reference standard, and their administration (ie, flow and timing). In addition concerns for applicability were assessed in the first three key areas.

For both risk of bias and concerns for applicability, the individual criteria were classified as 'low', 'high' or 'unclear' and the results were presented using tables available on the QUADAS website (http://www.quadas.org).

Detailed information on the target population was extracted according to the QUADAS II checklist concerning participants.[12]

Grading was applied to each study for rating the quality of evidence.[13]

Statistical analysis

Meta-analyses were conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines.[14] All quantitative data of selected studies were uniformed as OR as effect size (ES), with corresponding 95% CI, by Comprehensive Meta-Analysis (CMA) software V.2.2 (Biostat, Englewood, New Jersey, USA). Using CMA, a test for outliers was performed and studies with residual p value <0.05 were eliminated. Q and I^2 statistics were used to test the homogeneity among ES results. To calculate overall combined ES, CMA provides different meta-analytic models: in particular, if the assumption of heterogeneity has been identified the random effect model is used; otherwise, the fixed-model is adopted. Resulting ORs were presented as forest plots with the corresponding 95% CI. The Q statistic was also used to test the significance of moderators.

The Egger linear regression method (available in CMA) was used to estimate potential publication bias. If Egger's method showed a statistically significant bias (p values <0.05), the 'trim and fill' method was used to adjust ES for bias in a funnel plot. Briefly, the 'asymmetric' trials on the right side of the funnel (ie, trials that have no left side counterpart) were first located. These trials were removed ('trimmed') from the funnel, leaving a symmetric remainder from which the true centre of the funnel was estimated by the standard meta-analysis procedure. The 'trimmed' trials were then replaced and their missing counterparts imputed ('filled'): these were mirror images of 'trimmed' trials with the mirror axis placed at the pooled estimate. This allowed the calculation of an adjusted overall CI.

For studies with available binary data, we considered sensitivity and specificity as ES using the Meta-Analysis of Diagnostic and Screening Test (Meta-DiSc) program, V.1.4 (freeware).[15] In addition, the estimate of SROC to describe the relationship between test sensitivity and specificity across all studies was considered.[15]

Data were presented as forest plots with the corresponding 95% CI. Positive (LR+) and negative (LR–) likelihood ratios, corresponding to sensitivity/(1−specificity) and (1−sensitivity)/specificity,[16] were also estimated and meta-analysed with Meta-DiSc. In particular, the strength of the indication for the presence of the disease provided by the positive result of the test is relevant when $LR+\geq 10$, modest when $5\leq LR+<10$, and poor when $2\leq LR+<5$, and the strength of the indication for the absence of the disease provided by the negative result of the test is relevant when $LR-\leq 0.10$, modest when $0.10<LR-\leq 0.20$, and poor when $0.20<LR-\leq 0.50$.[17] Finally, for each study, positive and negative predictive values were estimated.

RESULTS

Features of retrieved studies

The search strategy retrieved a total of 252 potentially eligible papers, restricted to 161 after removing duplicate records. After evaluation of titles and abstracts, a further 106 records were excluded and a total of 55 original articles were preliminary considered eligible for the full text examination. Among those, 39 papers were excluded because of:

- sensitivity and decisional threshold for HE4/CA-125 were estimated by fixing specificity (n=20);
- diagnostic sensitivity and specificity were obtained only using a diagnostic algorithm in which HE4 and/or CA-125 was included (n=3);
- partial or total inclusion in the control group of healthy individuals (n=10);

- only median HE4/CA-125 concentrations were available (n=4); and
- reported markers evaluation was on healthy subjects only (n=2).

Finally, a total of 16 articles met the criteria to be included in the meta-analysis (see online supplementary figure S1).[18–33] The main characteristics of selected studies are summarised in table 1, and table 2 shows data from studies with binary data presentation, including the prevalence of OC, the adopted cut-off, and parameters related to the diagnostic performances of HE4 and CA-125 in each study.

Population

In all studies, participants were enrolled because of the presence of gynaecological disease or, more specifically, of a pelvic mass suspected for OC.[22 26–31 33] According to the QUADAS II checklist, details of the selection criteria, enrolment, sampling and data collection were retrieved.

Only a few studies adopted restrictive selection criteria by excluding pregnant women,[22 28] subjects with presence or previous history of cancer,[19 22 28] oophorectomy[26] or positive to breast cancer gene expression.[32] One study included in the case group patients with low-malignant tumours potentially not detectable by biomarkers and another included in the control group only women with endometriosis.[23 30] As evidenced in table 1, the enrolment widely differed across studies for the following:

1. Setting of data collection (gynaecology–oncology or gynaecology).
2. Sample size and OC prevalence.
3. Patient characteristics (ie, prevalence of women of postmenopausal status).
4. Severity of OC (ie, prevalence of late stages).

Each of these points theoretically represented a source of heterogeneity among studies likely influencing the pre-test disease probability. Forty-four per cent of the studies were performed in a gynaecology–oncology setting, suggesting a different assessment of the disease and a higher grade of severity for OC. It is noteworthy that studies including early OC stages were performed in gynaecology. As table 1 and table 2 clearly show, the sample size and OC prevalence widely influenced the precision of the estimates and the reliability of markers' diagnostic parameters. Wide differences in the prevalence of women in postmenopausal status across studies should influence HE4 and CA-125 diagnostic performances. Similarly discrepant performances may be reported according to the prevalence of early OC stages.

In most studies the recruitment was based on the result of transvaginal ultrasonography[18 19 22–27 31 32]; in the remainder it was according to clinical and laboratory data. Both prospective enrolment of patients and retrospective collection of data were performed to assure the evaluation of continuous case series.

Specimen collection was quite similar for all studies: venous blood was generally drawn before surgery into tubes containing no anticoagulants (EDTA was used in only one study)[25] and, after centrifugation, samples were stored at −70/−80°C until measurements were done. Most studies used the manual HE4 enzyme immunoassay from Fujirebio Diagnostic (n=7) or the fully automated chemiluminescent microparticle-based assay on

Table 1 Main characteristics of selected studies

Study no. (ref)	Patients no. (OC vs BGD)	Company/platform of HE4 assay	Company/platform of CA-125 assay	Enrolling centre	Study design	Data presentation
1 ([18])	66 vs 257	Abbott/Architect	Abbott/Architect	G	CS	Binary data
2 ([19])	113* vs 165†	Abbott/Architect	Abbott/Architect	GO	CS	Binary data
3 ([20])	125‡ vs 289	Abbott/Architect	Abbott/Architect	G	CC	Binary data
4 ([21])	111§ vs 285¶	Abbott/Architect	Abbott/Architect	G	CS	Binary data
5 ([22])	34** vs 195	Abbott/Architect	Abbott/Architect	G	PCT	Binary data
6 ([23])	52 vs 150	CanAg ELISA	CanAg ELISA	GO	CS	Binary data
7 ([24])	96 vs 90	Abbott/Architect—Fujirebio ELISA	Abbott/Architect—Fujirebio ELISA	GO	CS	Means
8 ([25])	29 vs 71	Fujirebio ELISA	Fujirebio ELISA	GO	CS	Binary data
9 ([26])	161†† vs 228‡‡	Fujirebio ELISA	Fujirebio ELISA	GO	PCT	Binary data
10 ([27])	55 vs 49	Fujirebio ELISA	Fujirebio ELISA	G	CC	Binary data
11 ([28])	149 vs 350	Luminex Multiplexed ELISA	Luminex Multiplexed ELISA	GO	PCT	Binary data
12 ([29])	37 vs 50	CanAg ELISA	CanAg ELISA	G	CS	Binary data
13 ([30])	41§§ vs 24	Fujirebio ELISA	Roche Diagnostics/Cobas e411	G	CC	Binary data
14 ([31])	32 vs 86	Fujirebio ELISA	Radim RIA	GO	CC	Binary data
15 ([32])	227¶¶ vs 158***	Luminex Multiplexed ELISA	Luminex Multiplexed ELISA	G	CS	Binary data
16 ([33])	14 vs 69†††	Fujirebio ELISA	Fujirebio ELISA	G	CS	Means

*26 pre- and 87 post-menopause.
†69 pre- and 96 post-menopause.
‡24 early (International Federation of Gynaecology and Obstetrics (FIGO) stage I/II) and 101 late (FIGO stage III/IV).
§27 pre- and 84 post-menopause.
¶226 pre- and 59 post-menopause.
**Pre-menopausal women only.
††42 pre- and 119 post-menopause.
‡‡142 pre- and 86 post-menopause.
§§13 early (FIGO stage I/II) and 28 late (FIGO stage III/IV).
¶¶58 pre- and 169 post-menopause (63 early (FIGO stage I/II) and 106 late (FIGO stage III/IV)).
***18 pre- and 140 post-menopause.
†††Ovarian endometriosis only.
BGD, benign gynaecological disease; CC, case–control study; CS, cross-sectional study; OC, ovarian cancer; G, gynaecology; GO, gynaecology–oncology; HE4, human epididymis protein 4; PCT, prospective clinical trial; RIA, radioimmunoassay.

Table 2 Diagnostic performance of human epididymis protein 4 (HE4) and carbohydrate antigen 125 (CA-125) in the subset of studies using binary data presentation

		HE4							CA-125						
Study no. (ref)	OC prevalence	Threshold (pmol/l)	Sens % (95% CI)	Spec % (95% CI)	LR+ (95% CI)	LR– (95% CI)	PPV %	NPV %	Threshold (kU/l)	Sens % (95% CI)	Spec % (95% CI)	LR+ (95% CI)	LR– (95% CI)	PPV %	NPV %
1 ([18])	0.20	33.2	90.9 (81 to 97)	94.2 (91 to 97)	15.2 (9.2 to 24.9)	0.10 (0.04 to 0.21)	80.0	97.6	38.3	72.7 (60 to 83)	94.6 (91 to 97)	13.0 (7.6 to 22.1)	0.29 (0.19 to 0.43)	77.4	93.1
2 ([19])	0.40	70*–140†	79.6 (71 to 87)	97.0 (93 to 99)	26.5 (11.1 to 63.2)	0.21 (0.15 to 0.30)	94.7	87.4	35	93.8 (88 to 96)	72.7 (65 to 79)	3.4 (2.7 to 4.4)	0.09 (0.04 to 0.18)	70.2	94.5
3 ([20])	0.30	140	75.2 (67 to 83)	98.6 (96 to 100)	53.6 (20.1 to 142.5)	0.25 (0.19 to 0.34)	95.9	90.2	35	80.0 (72 to 87)	67.1 (61 to 73)	2.4 (2.0 to 2.9)	0.30 (0.21 to 0.43)	51.3	88.6
4 ([21])	0.28	150	79.3 (71 to 86)	98.9 (97 to 100)	79.0 (25.5 to 244.5)	0.21 (0.15 to 0.31)	96.7	92.5	35	82.9 (75 to 89)	70.9 (65 to 76)	2.9 (2.3 to 3.5)	0.24 (0.16 to 0.36)	52.6	91.4
5 ([22])	0.15	70	64.7 (47 to 80)	91.8 (87 to 95)	8.1 (64.8 to 13.8)	0.38 (0.24 to 0.61)	57.9	93.7	35	85.3 (69 to 95)	59.5 (52 to 66)	2.1 (1.7 to 2.6)	0.25 (0.11 to 0.56)	26.9	95.9
6 ([23])	0.26	150	73.1 (59 to 84)	98.7 (95 to 100)	73 (18.2 to 192.1)	0.27 (0.17 to 0.43)	95.0	91.4	35	88.5 (77 to 96)	58.0 (50 to 66)	2.1 (1.7 to 2.6)	0.21 (0.10 to 0.44)	42.2	93.5
8 ([25])	0.29	70	86.2 (68 to 96)	85.9 (76 to 93)	6.1 (3.4 to 11.1)	0.16 (0.07 to 0.41)	71.4	93.8	35	69.0 (49 to 85)	90.1 (81 to 96)	6.9 (3.3 to 14.5)	0.34 (0.20 to 0.60)	74.1	87.7
9 ([26])	0.41	70*–150†	74.5 (67 to 81)	83.3 (78 to 88)	4.5 (3.3 to 6.1)	0.31 (0.23 to 0.40)	75.9	82.3	35	79.5 (72 to 86)	81.6 (76 to 86)	4.3 (3.3 to 5.7)	0.25 (0.18 to 0.34)	75.3	84.9
10 ([27])	0.53	74.2	76.4 (63 to 87)	93.9 (83 to 99)	10.9 (3.6 to 33.0)	0.25 (0.16 to 0.41)	93.3	78.0	35	70.9 (57 to 82)	77.6 (63 to 88)	3.2 (1.8 to 5.5)	0.38 (0.24 to 0.58)	78.0	70.4
11 ([28])	0.30	54.8	87.2 (81 to 92)	89.4 (86 to 92)	8.2 (6.0 to 11.2)	0.15 (0.10 to 0.22)	77.8	94.3	52.5	74.5 (67 to 81)	93.7 (91 to 96)	11.8 (7.8 to 17.9)	0.27 (0.21 to 0.36)	83.5	89.6
12 ([29])‡	0.43	25	100 (91 to 100)	100 (93 to 100)	–	–	100	100	37	83.8 (68 to 94)	100 (93 to 100)	–	0.15 (0.07 to 0.31)	100	89.3
13 ([30])	0.63	72	82.9 (68 to 93)	87.5 (68 to 97)	6.6 (2.3 to 19.3)	0.20 (0.10 to 0.39)	91.9	75.0	35	73.2 (57 to 86)	79.2 (58 to 93)	3.5 (1.6 to 7.8)	0.34 (0.20 to 0.58)	85.7	63.3
14 ([31])§	0.27	150	96.9 (84 to 100)	100 (96 to 100)	–	0.03 (0.00 to 0.21)	100	98.9	35	87.5 (71 to 97)	88.4 (80 to 94)	7.5 (4.2 to 13.7)	0.14 (0.06 to 0.35)	73.3	95.0
15 ([32])	0.59	73.7¶	73.1 (67 to 79)	85.4 (79 to 91)	5.0 (3.4 to 7.4)	0.31 (0.25 to 0.39)	87.8	68.9	93.15	74.9 (69 to 80)	82.9 (76 to 88)	4.4 (3.1 to 6.3)	0.30 (0.24 to 038)	86.3	69.7

*In pre-menopausal women.
†In post-menopausal women.
‡Outlier for CA-125.
§Outlier for HE4.
¶Original cut-off in pg/l, converted in pmol/l by using a multiplication factor of 0.04.
LR+, positive likelihood ratio; LR–, negative likelihood ratio; NPV, negative predictive value; OC, ovarian cancer; PPV, positive predictive value; Sens, sensitivity; Spec, specificity.

the Abbott Achitect platform (n=6). For these assays, employing the same antibodies, but differing with respect to the signal detection technology (colorimetry vs chemiluminescence), similar diagnostic performance has been reported in a head-to-head comparison by Ruggeri *et al*.[24]

As reported in the table 2, decision thresholds for HE4 differed across studies, even in those employing the same assay. Most studies seem to adopt decision thresholds suggested by the assay manufacturer, differentiated for menopausal status only for the Architect assay. Others derived threshold values from a parallel cohort of healthy controls or from a training group,[18 27 29 32] or by maximising sensitivity and specificity.[28 30]

Quality and level of evidence from individual studies

The risk of bias for patient selection, index test, reference standard, flow and timing as well as the concerns for applicability related to the first three domains are shown in figure 1.

All the selected studies avoided spectrum bias by evaluating the diagnostic performances of HE4 and CA-125 in women with a gynaecological disease and suspected as having OC, thus meeting the aim of the review. Notably, eight studies selected only women with a pelvic mass,[22 26–31 33] who might be the most appropriate population to be submitted to evaluation of circulating markers for diagnostic purposes. With only two exceptions,[20 21] studies assured a consecutive enrolment, thus avoiding a selection bias. In 50% of studies a partial verification bias may occur since not all patients with benign gynaecological disease were diagnosed with the reference diagnostic method.[22 26–31 33] Most studies retrospectively enrolled patients with available clinical data, and had a cross-sectional design (n=9) or were case–control studies (n=4), whereas only three articles were prospective clinical trials (PCTs).

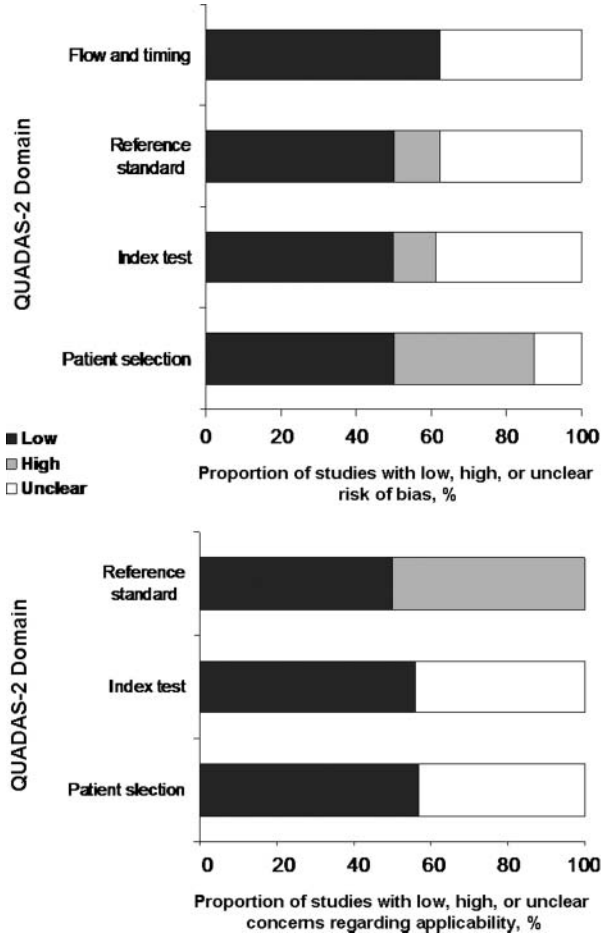

Figure 1 Graphical display of study characteristics according to QUADAS II recommendations.

The score related to the risk of bias and applicability for the conduct and interpretation of the index test mainly accounted for the use of assays with likely different performances (25% of studies), and for the selection of the diagnostic thresholds. Notably, in a minority of studies an overestimation of diagnostic accuracy may be suspected because of the application of data-driven cut-off values (ie, the best threshold).[28 30]

The greatest concern in the category of applicability was related to patient selection. Fifty per cent of the studies strictly selected patients with a well characterised pelvic mass and did not consider the wider framework of gynaecological diseases. In addition, only on these selected patients there are no concerns on the applicability of the reference standard. Concerns about the applicability of the index test might be overcome by working on the commutability of assays and on their diagnostic thresholds.

Finally eight studies were suggested to provide quite reliable evidence,[22 26–31 33] as they fulfilled the QUADAS II requirements for good quality research.

However, resorting to GRADE guidelines, we were able to rate the quality of the body of evidence. Only three (19%) of the studies provided high quality and level of evidence; most evidence was classified as low for both domains.

This implies that our confidence in the effect of estimate is quite limited. There might be a not negligible risk that the true diagnostic performance might be overestimated.

Meta-analysis results

Among the 16 studies included in the meta-analysis, two for HE4 and one for CA-125 were identified as outliers and thus eliminated.[29 31] Figure 2 shows the random overall combined ES shown as a forest plot of the OR and corresponding 95% CI. The OR was significant for both HE4 (43.2, 95% CI 21.9 to 85.4) and CA-125 (15.4, 95% CI 10.4 to 22.8) in a heterogeneous set of studies (for HE4: Q=117.2, p<0.0001, I^2=88.9%; for CA-125: Q=60.0, p<0.0001, I^2=76.6%). OC prevalence, type of assays, type of enrolling medical centre, and study design were analysed as moderators, but none of these characteristics influenced total ES. The Egger linear regression showed a significant publication bias (p=0.03) only for HE4 outcome. Using the 'trim and fill' method, the adjusted value of the overall combined ES was 37.2 (95% CI 19.0 to 72.7) with one trimmed study (see online supplementary figure S2).

Among the 14 studies displaying binary data, one for HE4 and one for CA-125 were identified as outliers and eliminated.[29 31] Meta-analyses revealed an overall sensitivity of 79% (95% CI 76% to 81%) and a specificity of 93% (95% CI 92% to 94%) for HE4, and an overall sensitivity of 79% (95% CI 77% to 82%) and a specificity of 78% (95% CI 76% to 80%) for CA-125 (figure 3). For these studies, the global OR was 61.1 (95% CI 31.5 to 118.5, adjusted for publication bias) for HE4 and 17.4 (95% CI 11.9 to 25.4) for CA-125.

Because of the heterogeneity of studies for both HE4 and CA-125, asymmetric SROC curves were used (see online supplementary figure S3). However, the visual inspection of the fitted curves revealed a wide scatter, in particular for HE4 studies, mainly due to a deceptive interpolation of a few points located in the upper left side of the graph. This is likely consequent to the similar sensitivity and specificity of marker assays, mainly

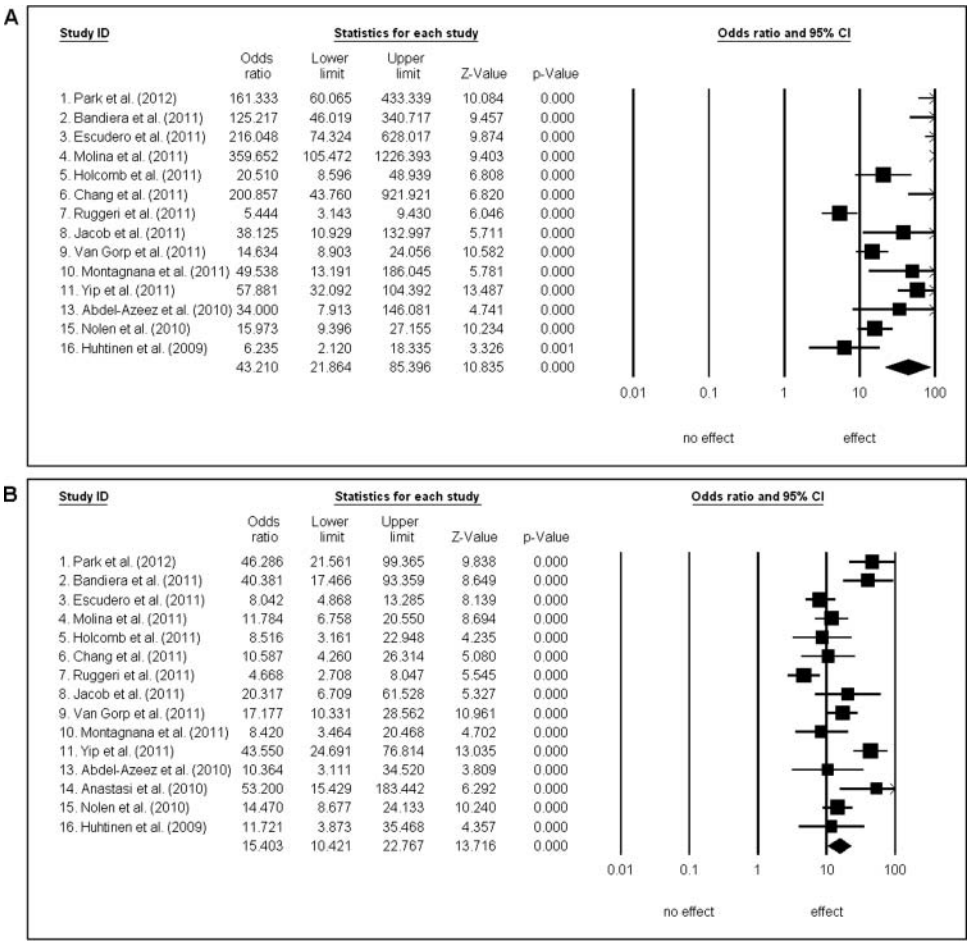

Figure 2 Random overall combined effect size of human epididymis protein 4 (HE4) (A) and carbohydrate antigen 125 (CA-125) (B) shown as forest plots of the OR with 95% CI.

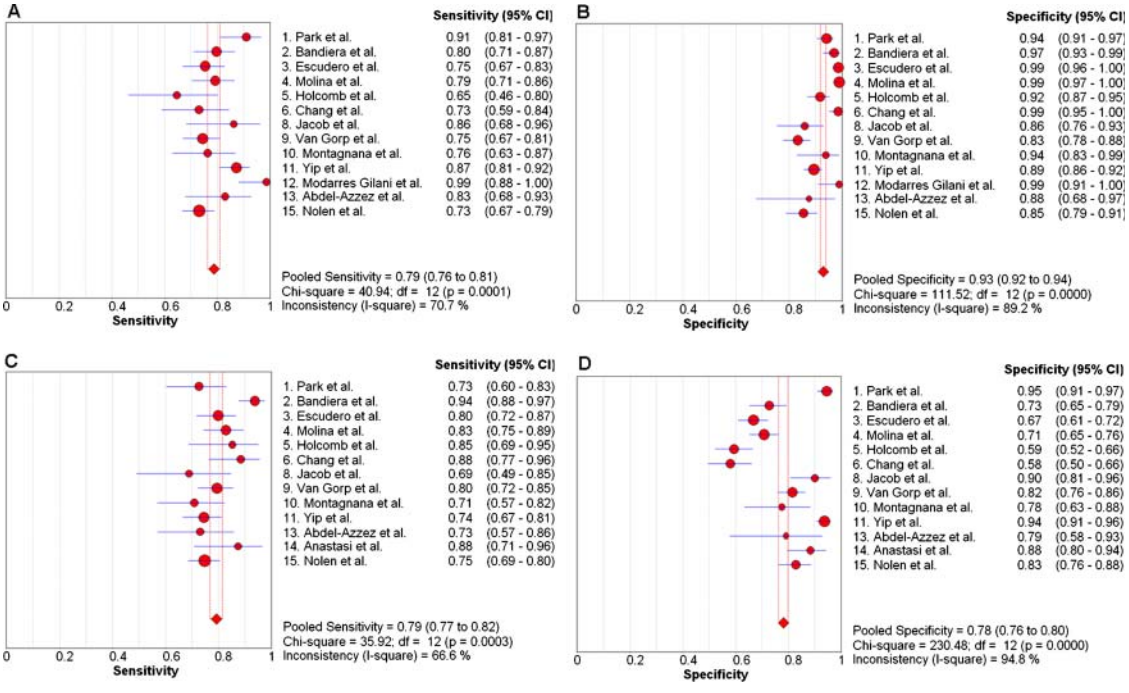

Figure 3 Sensitivity and specificity plots of human epididymis protein 4 (HE4) (A and B, respectively) and carbohydrate antigen 125 (CA-125) (C and D, respectively) determination in the diagnosis of ovarian cancer.

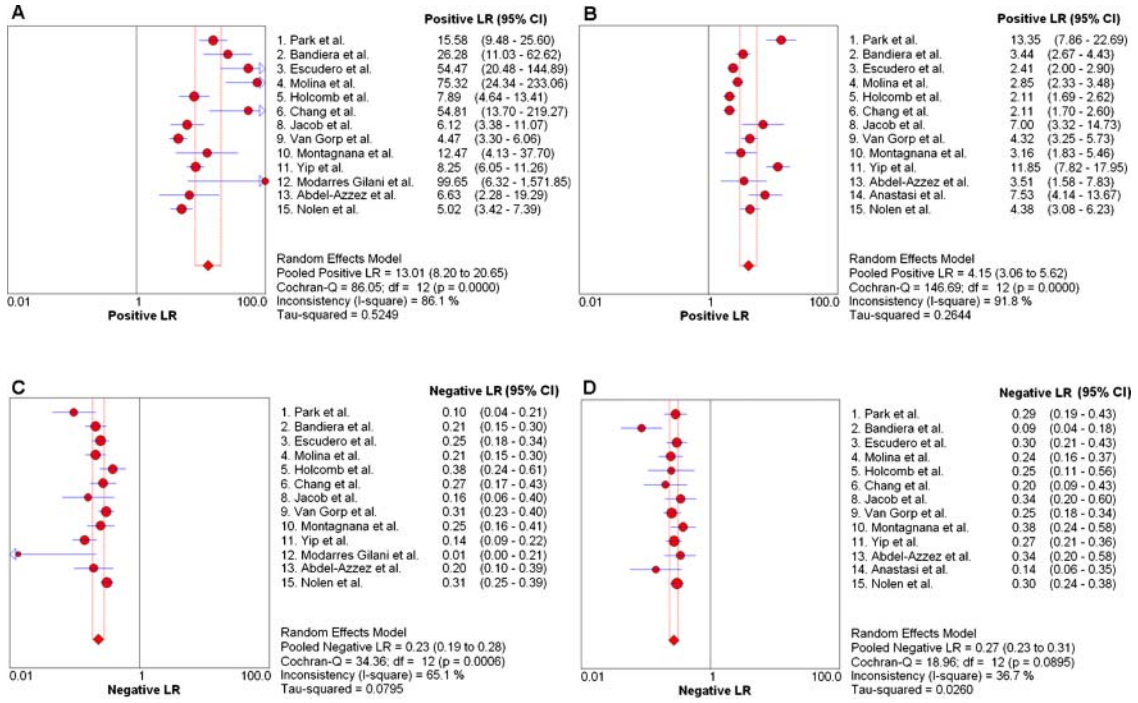

Figure 4 Likelihood ratio plots of human epididymis protein 4 (HE4) (A and C) and carbohydrate antigen 125 (CA-125) (B and D) determination in the diagnosis of ovarian cancer.

due to the use of the same antibodies and to overlapping/harmonised thresholds.

LR+ and LR– were 13.0 (95% CI 8.2 to 20.7) and 0.23 (95% CI 0.19 to 0.28) for HE4, and 4.2 (95% CI 3.1 to 5.6) and 0.27 (95% CI 0.23 to 0.31) for CA-125 (figure 4). Knowledge of LR+ is mandatory to evaluate the capability of the marker to recognise OC in suspected women. In this regard, our meta-analysis showed a higher LR+ for HE4 than for CA-125, assigning to HE4 a relevant capability for ruling OC in. On the contrary, both markers displayed relatively high LR–, indicating a poor capability to exclude the presence of OC. Given the high heterogeneity of enrolled patient populations in different studies, a meta-analysis of predictive values was not done.

An additional evaluation of diagnostic performance of the combined measurements of HE4 and CA-125 was performed in four studies.[22 23 30 32] Their meta-analysis resulted in a pooled sensitivity of 82% (95% CI 78% to 86%) and a pooled specificity of 76% (95% CI 72% to 80%).

Four studies specifically evaluated HE4 and CA-125 on subgroups of post-menopausal women.[19 21 26 32] However, the small study size (a total of 459 cases vs 381 controls) and the huge heterogeneity among studies did not permit any data pooling and further statistical elaboration. Due to a few articles evaluating biomarker performance in the detection of OC at an early stage (International Federation of Gynaecology and Obstetrics (FIGO) stages I and II),[20 30 32] this issue also needs to be further studied.

DISCUSSION

The main challenge for laboratory biomarkers of OC diagnosis is to allow the accurate detection of malignancy as early as possible to improve clinical outcome and survival of patients.[34] Currently, CA-125 is the most widely used marker in OC diagnostics, even if there is no agreement among different guidelines on the use of CA-125 for the screening and evaluation of high-risk women in a primary care setting.[34–36] This is possibly because supporting evidence is only indirect, coming from systematic reviews of studies performed in secondary care settings that may significantly differ in case mix.[3] On the other hand, the intrinsic limitations of CA-125 have greatly stimulated the search of additional biomarkers sought to improve the accuracy for identifying malignancy in women with a pelvic mass. Among others, HE4 has been reported as the most promising marker to aid in OC diagnosis.[9] The only available meta-analysis evaluating its diagnostic value is, however, affected by important methodological limitations.[11] First, the study failed to evaluate the HE4 diagnostic performance in the right clinical context (ie, women with a suspected gynaecological disease): Yu *et al* did not exclude studies partially or totally enrolling in the control group healthy subjects, a clinically not relevant population, with a possible spurious increase in the clinical efficacy of the marker.[11 37] Second, the use of symmetric SROC curves in synthesising diagnostic accuracy is prone to criticisms whether included studies display a wide heterogeneity as in the case of HE4 and CA-125. In these conditions, it is recommended to adopt asymmetric SROC curves. Even in this case, at visual inspection SROC curves may appear not appropriate, as in the case of CA-125 and HE4: the model is indeed underpowered to obtain reliable estimates with adequate precision.[38] The obtained asymmetric SROC curves were fitted according to the lowest number of points covering a tight area of the graph; thus, it seems unreasonable to obtain a curve reflecting test performances at highest sensitivity and specificity, where no points were observed (lower left and upper right corner of the graph). Such a situation may occur when most studies resort to the same assay and/or to similar thresholds or results restricted to a narrow range of values. The wide imprecision of the obtained SROC curves as well as the fact that they may result from a deceptive and thus misleading mathematical interpolation can make their interpretation mistaken. Furthermore, the review by Yu *et al* did not deal with some relevant clinical questions

concerning the introduction of HE4, such as its effectiveness in post-menopausal women and in early-stage OC. Finally, the evaluation of diagnostic performance of the combined measurements of HE4 and CA-125 was not considered, although this was the original intended application of the marker.[10]

Our results showed that women with gynaecological disease and increased concentrations of HE4 or CA-125 are at higher risk for malignancy. In particular, the risk for OC is significantly increased for patients with HE4 positive results (OR 37.2). As expected from immunohistochemical data,[10] the sensitivity of HE4 and CA-125 overlapped (79%), while HE4 exhibited a significantly higher specificity than CA-125 (93% vs 78%). The LR calculation confirmed that HE4 outperforms CA-125 in identifying OC (LR+: 13.0 vs 4.2), whereas the capability to rule out OC was quite similar for both markers and rather poor. These results support the hypothesis that HE4 could replace CA-125 measurement as a standalone biochemical test for OC diagnosis more than improve its diagnostic performance by combining their measurements.

Although the evidence of diagnostic effectiveness in detecting early-stage tumours in post-menopausal women is of pivotal relevance, there are currently not enough studies for estimating HE4 performance in this clinical scenario. In particular, the focus on menopausal status is of relevance since guidelines assign the highest baseline risk index to post-menopausal women.[34] The possibility that HE4 may differently perform according to the menopausal status is not marginal since higher HE4 concentrations are physiologically detectable in post-menopausal women and this may require the definition of specific clinical thresholds for this condition.[34]

Limitations to the clinical validity of presented results are the significant publication bias for HE4 studies and the heterogeneity among retrieved studies. The adoption of the 'trim and fill' method to adjust pooled estimates for the publication bias suggested, however, a relatively marginal effect of this limiting condition and the recalculated OR was not significantly lower than the unadjusted one.[39] The estimated heterogeneity among studies initially seemed mainly due to different sample size leading to a slight funnel plot asymmetry. However, additional sources of heterogeneity emerged by evaluating the quality of primary studies that often resulted in suboptimal resorting to QUADAS checklist. Differences in study design, in clinical source of patients, in the adoption of eligibility criteria (eg, inclusion of patients with low-malignant potential ovarian tumours) were evidenced. In addition, studies often did not look comparable for the uneven distribution of patients in pre- and post-menopausal status, OC histological subtypes, and OC FIGO stages. Furthermore, only in some studies was the OC diagnosis performed by or in collaboration with a gynaecologist oncologist. This is not a marginal issue as there is a lively debate about the different prognostic impact of diagnostic management when performed by surgeons with an appropriate expertise or by the gynaecologist alone.[40 41] Another relevant issue was represented by the HE4 concentration adopted as decision threshold for OC diagnosis, since across studies there was often a subjective adoption of different threshold levels, even when using the same assay. Finally, there was no agreement about the need to select different HE4 thresholds for pre- and post-menopausal women.

In conclusion, there is only a preliminary and mild evidence on the ability of HE4 measurement in serum to overcome CA-125 in terms of diagnostic performance for identification of OC in women with suspected gynaecological disease. The suboptimal quality of research, the modest level of evidence and the lack of agreement on decision thresholds are likely to hamper the clinical value of the marker. Before integrating HE4 in the OC diagnostic algorithm, in order to replace or to complete the CA-125 information, more robust estimates of HE4 diagnostic performances are needed. In particular, well designed PCTs are required to reinforce this preliminary evidence and, in particular, to evaluate the HE4 capability to identify OC at early stage in post-menopausal women with a pelvic mass.

Key messages

- The risk for OC is significantly increased for patients with HE4 positive results (OR 37.2).
- HE4 exhibited a significantly higher specificity than CA-125 (93% vs. 78%).
- HE4 outperforms CA-125 in identifying OC (LR+: 13.0 vs. 4.2).

Contributors SF and FB initiated and supervised the study, performed the analysis and wrote the paper. MP contributed to the design of the study, critically revised the paper and introduced relevant scientific remarks. ML helped collect the published articles. ML, EB and PB revised the statistical approach. All of the authors have read and approved the final paper.

Competing interests None.

Patient consent Obtained.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1 Ferlay J, Shin HR, Bray F, *et al*. *GLOBOCAN 2008, cancer incidence and mortality worldwide: International Agency for Research on Cancer*. Lyon, France, 2010. http://globocan.iarc.fr (accessed May 2012).
2 Fountain J, Trimble E, Birrer MJ. Summary and discussion of session recommendations. *Gynecol Oncol* 2006;103:S23–25.
3 National Institute for Health and Clinical Excellence (NICE): Guidance. *Ovarian cancer: the recognition and initial management of ovarian cancer. National Collaborating Centre for Cancer (UK)*. Cardiff, UK: National Collaborating Centre for Cancer, 2011.
4 Sturgeon CM, Duffy MJ, Stenman UH, *et al*. National Academy of Clinical Biochemistry laboratory medicine practice guidelines for use of tumor markers in testicular, prostate, colorectal, breast, and ovarian cancers. *Clin Chem* 2008;54:e11–79.
5 Bast RC Jr. Status of tumor markers in ovarian cancer screening. *J Clin Oncol* 2003;21:S200–205.
6 Köbel M, Kalloger SE, Boyd N, *et al*. Ovarian carcinoma subtypes are different diseases: implications for biomarker studies. *PLoS Med* 2008;5:1749–60.
7 Miralles C, Orea M, España P, *et al*. Cancer antigen 125 associated with multiple benign and malignant pathologies. *Ann Surg Oncol* 2003;10:150–4.
8 Havrilesky LJ, Whitehead CM, Rubatt JM, *et al*. Evaluation of biomarker panels for early stage ovarian cancer detection and monitoring for disease recurrence. *Gynecol Oncol* 2008;110:374–82.
9 Hellström I, Raycraft J, Hayden-Ledbetter M, *et al*. The HE4 (WFDC2) protein is a biomarker for ovarian carcinoma. *Cancer Res* 2003;63:3695–700.
10 Rosen DG, Wang L, Atkinson JN, *et al*. Potential markers that complement expression of CA-125 in epithelial ovarian cancer. *Gynecol Oncol* 2005;99:267–77.
11 Shuang Y, Yang H, Xie S, *et al*. Diagnostic value of HE4 for ovarian cancer: a meta-analysis. *Clin Chem Lab Med* 2012;50:1439–46.
12 Whiting PF, Rutjes AW, Westwood ME, *et al*. QUADAS-2 Group. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med* 2011;155:529–36.
13 Balshem H, Helfand M, Schünemann HJ, *et al*. GRADE guidelines: 3. Rating the quality of evidence. *J Clin Epidemiol* 2011;64:401–6.
14 Liberati A, Altman DG, Tetzlaff J, *et al*. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *PLoS Medicine* 2009;6:e1000100.
15 Zamora J, Abraira V, Muriel A, *et al*. Meta-DiSc: a software for meta-analysis of test accuracy data. *BMC Med Res Methodol* 2006;6:31.
16 Deeks JJ, Altman DG. Diagnostic test: likelihood ratios. *Br Med J* 2004;329:168–9.

17 Pepe MS, Feng Z, Janes H, *et al*. Pivotal evaluation of the accuracy of a biomarker used for classification or prediction: standards for study design. *J Natl Cancer Inst* 2008;100:1432–8.
18 Park Y, Kim Y, Lee EY, *et al*. Reference ranges for HE4 and CA-125 in a large Asian population by automated assays and diagnostic performances for ovarian cancer. *Int J Cancer* 2012;130:1136–44.
19 Bandiera E, Romani C, Specchia C, *et al*. Serum human epididymis protein 4 and risk for ovarian malignancy algorithm as new diagnostic and prognostic tools for epithelial ovarian cancer management. *Cancer Epidemiol Biomarkers Prev* 2011;20:2496–506.
20 Escudero JM, Auge JM, Filella X, *et al*. Comparison of serum human epididymis protein 4 with cancer antigen 125 as a tumor marker in patients with malignant and nonmalignant diseases. *Clin Chem* 2011;57:1534–44.
21 Molina R, Escudero JM, Augé JM, *et al*. HE4 a novel tumour marker for ovarian cancer: comparison with CA 125 and ROMA algorithm in patients with gynaecological diseases. *Tumour Biol* 2011;32:1087–95.
22 Holcomb K, Vucetic Z, Miller MC, *et al*. Human epididymis protein 4 offers superior specificity in the differentiation of benign and malignant adnexal masses in premenopausal women. *Am J Obstet Gynecol* 2011;205:358.e1–6.
23 Chang X, Ye X, Dong L, *et al*. Human epididymis protein 4 (HE4) as a serum tumor biomarker in patients with ovarian carcinoma. *Int J Gynecol Cancer* 2011;21:852–8.
24 Ruggeri G, Bandiera E, Zanotti L, *et al*. HE4 and epithelial ovarian cancer: comparison and clinical evaluation of two immunoassays and a combination algorithm. *Clin Chim Acta* 2011;412:1447–53.
25 Jacob F, Meier M, Caduff R, *et al*. No benefit from combining HE4 and CA-125 as ovarian tumor markers in a clinical setting. *Gynecol Oncol* 2011;121:487–91.
26 Van Gorp T, Cadron I, Despierre E, *et al*. HE4 and CA-125 as a diagnostic test in ovarian cancer: prospective validation of the Risk of Ovarian Malignancy Algorithm. *Br J Cancer* 2011;104:863–70.
27 Montagnana M, Danese E, Ruzzenente O, *et al*. The ROMA (Risk of Ovarian Malignancy Algorithm) for estimating the risk of epithelial ovarian cancer in women presenting with pelvic mass: is it really useful? *Clin Chem Lab Med* 2011;49:521–5.
28 Yip P, Chen TH, Seshaiah P, *et al*. Comprehensive serum profiling for the discovery of epithelial ovarian cancer biomarkers. *PLoS One* 2011;6:e29533.
29 Modarres-Gilani M, Ghaemmaghami F, Mousavi A, *et al*. Simultaneous measurement of two serum markers (CA-125 and HE-4) while diagnosing malignant ovarian epithelial tumors. *Pak J Med Sci* 2011;27:858–61.
30 Abdel-Azeez HA, Labib HA, Sharaf SM, *et al*. HE4 and mesothelin: novel biomarkers of ovarian carcinoma in patients with pelvic masses. *Asian Pac J Cancer Prev* 2010;11:111–16.
31 Anastasi E, Marchei GG, Viggiani V, *et al*. HE4: a new potential early biomarker for the recurrence of ovarian cancer. *Tumour Biol* 2010;31:113–19.
32 Nolen B, Velikokhatnaya L, Marrangoni A, *et al*. Serum biomarker panels for the discrimination of benign from malignant cases inpatients with an adnexal mass. *Gynecol Oncol* 2010;117:440–5.
33 Huhtinen K, Suvitie P, Hiissa J, *et al*. Serum HE4 concentration differentiates malignant ovarian tumours from ovarian endometriotic cysts. *Br J Cancer* 2009;100:1315–19.
34 Moore RG, Bast RC Jr. How do you distinguish a malignant pelvic mass from a benign pelvic mass? Imaging, biomarkers, or none of the above. *J Clin Oncol* 2007;25:4159–61.
35 American College of Obstetricians and Gynecologists (ACOG). ACOG Practice Bulletin. Management of adnexal masses. *Obstet Gynecol* 2007;110:201–14.
36 National Comprehensive Cancer Network (NCCN). *Clinical Practice Guidelines in Oncology. Ovarian cancer*. Version 2.2009. Washington, PA: NCCN, 2009.
37 Zweig MH, Robertson EA. Why we need better test evaluation. *Clin Chem* 1982;28:1272–6.
38 Chappell FM, Raab GM, Wardlaw JM. When are summary ROC curves appropriate for diagnostic meta-analyses? *Stat Med* 2009;28:2653–68.
39 Peters JL, Sutton AJ, Jones DR, *et al*. Performance of the trim and fill method in the presence of publication bias and between-study heterogeneity. *Stat Med* 2007;26:4544–62.
40 Earle CC, Schrag D, Neville BA, *et al*. Effect of surgeon specialty on processes of care and outcomes for ovarian cancer patients. *J Natl Cancer Inst* 2006;98:172–80.
41 McCluggage WG. Ten problematical issues identified by pathology review for multidisciplinary gynaecologicaloncology meetings. *J Clin Pathol* 2012;65:41–5.