Human Reproduction Update, Vol.17, No.3 pp. 301–310, 2011
Advanced Access publication on January 12, 2011 doi:10.1093/humupd/dmq060

human
reproduction
update

Chlamydia antibody testing and diagnosing tubal pathology in subfertile women: an individual patient data meta-analysis

K.A. Broeze [1,2,*], B.C. Opmeer [2], S.F.P.J. Coppus [1], N. Van Geloven [2], M.F.C. Alves [3], G. Ånestad [4], S. Bhattacharya [5], J. Allan [6], M.F. Guerra-Infante [7], J.E. Den Hartog [8], J.A. Land [9], A. Idahl [10], P.J.Q. Van der Linden [11], J.W. Mouton [12], E.H.Y. Ng [13], J.W. Van der Steeg [1], P. Steures [1], H.F. Svenstrup [14], A. Tiitinen [15], B. Toye [16], F. Van der Veen [1], and B.W. Mol [1,2]

[1]Centre for Reproductive Medicine, Department of Obstetrics and Gynaecology, Academic Medical Centre (AMC), Amsterdam, The Netherlands [2]Department of Clinical Epidemiology, Biostatistics and Bio-informatics, Academic Medical Centre (AMC), Amsterdam, The Netherlands [3]Department of Microbiology, Immunology, Parasitology, and Pathology, Institute of Tropical Pathology and Public Health, Goias, Brazil [4]Department of Virology, Norwegian Institute of Public Health, Oslo, Norway [5]Department of Obstetrics and Gynaecology, Aberdeen Maternity Hospital, Aberdeen, UK [6]Department of Reproductive Medicine, Wesley Hospital, Brisbane, Australia [7]Department Infectious Diseases, National Institute of Perinatology, Mexico City, Mexico [8]Department of Obstetrics and Gynaecology, Maastricht University Medical Center, Maastricht, The Netherlands [9]Department of Obstetrics and Gynaecology, University Medical Centre Groningen, Groningen, The Netherlands [10]Department of Clinical Sciences, Obstetrics and Gynaecology, Umeå University, Umeå, Sweden [11]Department of Obstetrics and Gynaecology, Deventer Hospital, Deventer, The Netherlands [12]Department of Medical Microbiology and Infectious Diseases, Canisius Wilhelmina, Hospital, Nijmegen, The Netherlands [13]Department of Obstetrics and Gynaecology, The University of Hong Kong, Hong Kong, China [14]Research Department of Reproductive Health, University College London, London, UK [15]Department of Obstetrics and Gynaecology, Helsinki University Central Hospital, Helsinki, Finland [16]Division of Microbiology, The Ottawa Hospital, University of Ottawa Hospital, Ottawa, Canada

*Correspondence address. E-mail: k.a.broeze@amc.uva.nl

Submitted on May 28, 2010; resubmitted on October 14, 2010; accepted on October 26, 2010

TABLE OF CONTENTS

BACKGROUND: The Chlamydia IgG antibody test (CAT) shows considerable variations in reported estimates of test accuracy, partly because of the use of different assays and cut-off values. The aim of this study was to reassess the accuracy of CAT in diagnosing tubal pathology by individual patient data (IPD) meta-analysis for three different CAT assays.

METHODS: We approached authors of primary studies that used micro-immunofluorescence tests (MIF), immunofluorescence tests (IF) or enzyme-linked immunosorbent assay tests (ELISA). Using the obtained IPD, we performed pooled receiver operator characteristics analysis and logistic regression analysis with a random effects model to compare the three assays. Tubal pathology was defined as either any tubal obstruction or bilateral tubal obstruction.

RESULTS: We acquired data of 14 primary studies containing data of 6191 women, of which data of 3453 women were available for analysis. The areas under the curve for ELISA, IF and MIF were 0.64, 0.65 and 0.75, respectively (P-value < 0.001) for any tubal pathology and 0.66, 0.66 and 0.77, respectively (P-value $= 0.01$) for bilateral tubal pathology.

CONCLUSIONS: In Chlamydia antibody testing, MIF is superior in the assessment of tubal pathology. In the initial screen for tubal pathology MIF should therefore be the test of first choice.

Key words: systematic review / individual patient data meta-analysis / chlamydia antibody test / tubal pathology

Introduction

Subfertility, defined as failure to conceive within 12 months despite regular unprotected intercourse, occurs in 10% of couples (Mosher and Pratt, 1991; Cahill and Wardle, 2002; Taylor, 2003). Besides ovulation disorders and sperm defects, tubal pathology is one of the main causes of subfertility. The prevalence of tubal pathology in subfertile couples ranges between 10 and 30% (Evers, 2002).

There are many tests to diagnose tubal pathology, of which Chlamydia IgG antibody test (CAT) testing, hysterosalpingography (HSG) and diagnostic laparoscopy (DLS) with chromopertubation are most often used (Broeze *et al.*, 2009). Although guidelines on the fertility work-up are not concordant, CAT is often recommended as a first-line test (British National Collaborating Center for Women's and Children's Health, 2004; Dutch Society of Obstetrics and Gynaecology, 2004). Unlike HSG or DLS, CAT is non-invasive and inexpensive (Mol *et al.*, 1997). When negative, CAT is thought to avoid unnecessary and invasive diagnostic testing, whereas in CAT positive women, further testing can be performed early, avoiding long-term expectative management (den Hartog *et al.*, 2008). A conventional meta-analysis reported that accuracy estimates of CAT for individual studies ranged between 21 and 90% for sensitivity and between 29 and 100% for specificity, with a summary receiver operator characteristics (ROC) estimating the accuracy of CAT to be moderate and comparable to that of HSG (Mol *et al.*, 1997).

Unfortunately, in the assessment of the diagnostic accuracy of CAT and the comparison of different assays over the full range of possible test results, conventional meta-analyses are impeded since they are based on the data reported at study level and limited to reported 2 × 2 tables. Pooling data to estimate overall accuracy is complicated by the use of different assays and different, mostly unreported, cut-off values in the included studies. Original data underlying the publications possibly contain CAT results recorded in a continuous rather than a dichotomized way, as well as information on several CAT assays within a single study and could be used to compare the accuracies of different tests (Broeze *et al.*, 2009).

The aim of this study was to assess the accuracy of several CAT assays in diagnosing tubal pathology, using continuous test results in an individual patient data (IPD) meta-analysis.

Methods

Literature search

In a previous meta-analysis on the accuracy of CAT, we identified all studies that compared CAT results to laparoscopy findings published until February 1996 (Mol *et al.*, 1997). An updated search was performed in Medline and Embase, from 1996 to 1 January 2010, using the words 'Chlamydia', '*Chlamydia trachomatis*', 'Chlamydia antibody', 'CAT' and 'tubal pathology', 'tubal infertility', 'tubal disease', 'fallopian tube disease', 'tubal obstruction' or 'tubal occlusion'. Cross-references of the selected articles were hand-searched and checked for other potentially eligible articles. No language restrictions were made. Two independent reviewers (K.A.B. and S.F.C.) screened the electronic searches for eligible articles by reading the title and abstract. Subsequently, we asked the authors of the selected articles to examine the provisional study list to identify any additional studies they may be aware of. In this way, data from studies that were missed by our search criteria, or data that were not published at all, were eligible for inclusion. We also considered inclusion of studies with potentially relevant data that were excluded in the previous conventional meta-analysis due to the inability to extract a 2 × 2 table.

Data collection

For each of the eligible articles, we obtained contact information on the first, second or last author through Medline, Embase or the internet. We approached these authors by email to inform them about the IPD meta-analysis project and invited them to share their data in this collaborative project. If authors were willing to participate, they were provided with a more detailed study proposal, and asked to send their original database. We requested the complete database in the original data format, to minimize the authors' input into going through their database to select the appropriate variables or to convert data into a specific format. Variables and categories needed to be adequately labeled within the dataset or in a separate data dictionary. Minimal requested data included the following variables: (anonymous) patient identifiers, CAT results and the results of laparoscopy (tubal pathology absent or present). CAT results of micro-immunofluorescence tests (MIF), immunofluorescence tests (IF) and *C. trachomatis* specific enzyme-linked immunosorbent assay tests (ELISA) were collected. These tests detect antibodies to *C. trachomatis* either by fluorescence of the elementary bodies of Chlamydia or by spectrophotometer detection of the complex of Chlamydia peptides and IgG antibodies (Wang and Grayston, 1970; Tuuminen *et al.*, 2000; Land *et al.*, 2003).

Tubal pathology was defined as either any tubal pathology or bilateral tubal pathology. Any tubal pathology was defined as at least unilateral occlusion of the fallopian tubes, with or without hydrosalpinges or peritubal adhesions. Bilateral tubal pathology was present when such abnormalities were seen in both tubes. The authors were asked to indicate whether tubal pathology was unilateral or bilateral, if possible. If authors had unpublished follow-up data available, they were asked to share these data as well. Ethical approval had been obtained by the original authors.

Quality assessment

The quality of the included studies was assessed according to the criteria of the QUADAS checklist (Whiting *et al.*, 2003). Additional items were created for the description of selection criteria, execution of tests and the diagnostic strategy that was used. Completeness of the data sets was described, based on the availability of data on patient identifiers, diagnostic test results and target disease. We compared the obtained data and the published results for consistency. Authors were contacted to confirm missing data or to check major discordant results. In addition, we organized a collaborators meeting, where authors clarified details of the original study designs and the tests performed. We used Review Manager (RevMan, Version 5.0. Copenhagen: The Nordic Cochrane Centre, the Cochrane Collaboration, 2008) to summarize the quality indicators according to QUADAS for the included studies.

Statistical analysis

We merged the data into a summary database if variables were compatible. Incompatible data were recoded before adding to the summary database. Data on MIF, IF and ELISA were analyzed separately. In studies that reported on assays from different manufacturers, we analyzed the most often used assay.

(i) First, we assessed the prevalences of both any tubal pathology and bilateral tubal pathology for the individual studies and for the complete set of included studies. We estimated heterogeneity in patient characteristics and prevalences between the included studies, using I^2 (Higgins *et al.*, 2003).

(ii) Second, we estimated sensitivity and specificity, based on the reported dichotomized CAT results per study.

(iii) Third, to correct for missing laparoscopies, thereby reducing verification bias, we performed multiple imputation per center. Verification bias can be defined as the occurrence of biased estimates when non-verified cases are omitted from the analyses. In the imputation step we assumed that, within one study, women that did not have a DLS had a tubal status comparable to the tubal status of women with similar CAT results, but who did have a DLS (Begg and Greenes, 1983; van der Heijden *et al.*, 2006; de Groot *et al.*, 2008). We performed five rounds of imputation, adjusted for center of origin and CAT result.

Subsequently, we performed a ROC analysis and calculated an area under the curve (AUC) for each individual study. The AUC was used to express the accuracy of a diagnostic test, representing the probability that a randomly chosen diseased patient is correctly ranked higher than a randomly chosen non-diseased patient. A perfect test, one with a perfect ROC curve, has an AUC of 1.0, whereas an uninformative test has an AUC of 0.5 (Hanley *et al.*, 1982).

Data from all the combined studies were analyzed in two ways. Overall accuracy of the CAT assays was first estimated by ROC-analysis on the empirical pooled data. In this analysis raw data were pooled for women with and without tubal pathology across studies and the resulting distributions were used to generate AUC estimates, without adjusting for the heterogeneity across studies. Then, the overall accuracy was estimated and ROC curves were generated by ROC-analysis based on the predicted probabilities from a random effects regression model. In the latter model we adjusted for the heterogeneity across studies, by including a random intercept that accounts for differences in prevalence of tubal pathology across studies, and by including a random effect which accounts for differences in test performance across studies. The random effects model assumes different possible values for the outcome of interest for each individual study and both variation within studies and between studies are taken into account (DerSimonian and Laird, 1986; Riley *et al.*, 2008). We used analysis of variance to test for significant differences in accuracy between the three assays. A *P*-value below 0.05 indicated statistical significance in these analyses.

Data were analyzed using SPSS 17.0 (SPSS Inc., Chicago, IL, USA) and SAS 9.1 (SAS Institute Inc., Cary, NC, USA).

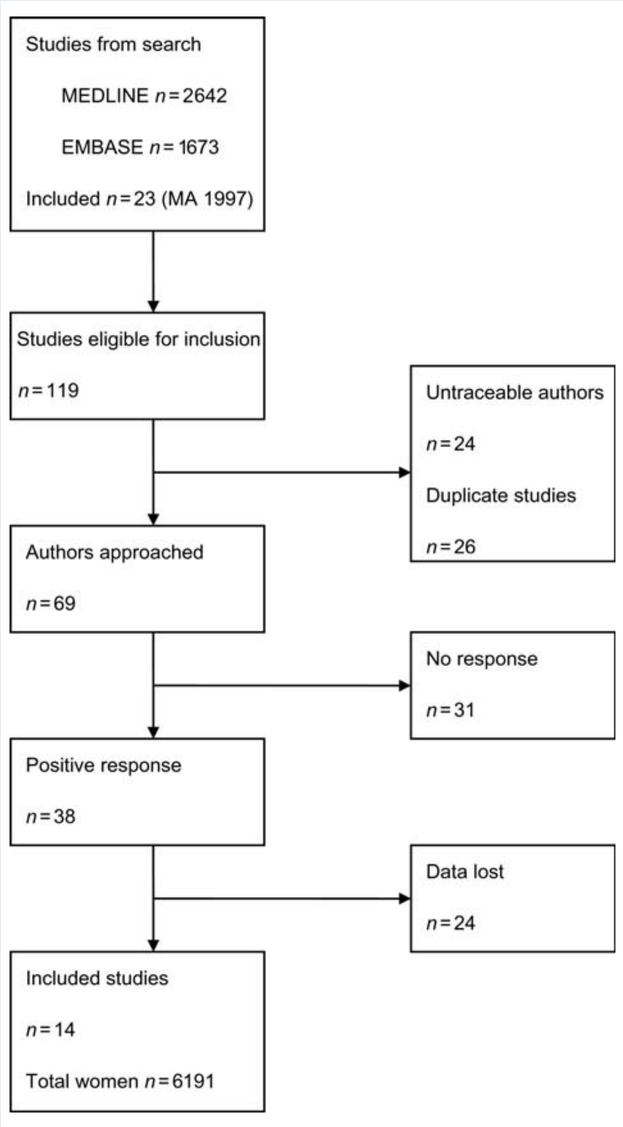

Figure 1 Flowchart of included studies.

Table I Overview of study characteristics of the included studies.

Study	Year	Total women	Inclusion criteria	Exclusion criteria	Study design	DLS performed	Diagnostic strategy
Bhattacharya	2003	207	Women referred for tubal evaluation	Previous tubal testing Previous tubal surgery	Cohort study (prospective)	Yes	CAT → DLS
Van der Steeg/ Steures	2007	3706	Women referred for subfertility evaluation	Previous tubal testing Previous tubal surgery	Cohort study (prospective) (multicenter)	Yes (depending on center)	CAT− → No TT CAT+ → HSG/DLS
Ng	2001	110	Women referred for tubal evaluation	Previous pelvic surgery Severe male factor	Cohort study (prospective)	Yes	CAT → DLS
Van der Linden	2002	395	Women referred for subfertility evaluation	Unknown	Cohort study (prospective)	Not standard	CAT− → HSG HSG+ → DLS HSG− → EXP → DLS
Allan	2004	52	Women referred for tubal evaluation	Ovulation disorders Severe male factor	Cohort study (prospective)	Yes	Unknown
Mouton	2002	107	Women referred for subfertility evaluation	Unknown	Cohort study (only subset of women were subfertility pts)	Not standard	Unknown
Land/Den Hartog	2005	662	Women referred for subfertility evaluation	Previous pelvic surgery	Cohort study (prospective)	Yes	CAT− → HSG CAT+ → DLS HSG+ → DLS HSG− → EXP → DLS
Idahl	2004	244	Women referred for subfertility evaluation	Unknown	Cohort study (prospective)	Not standard	CAT → HSS or DLS HSS+ → DLS HSS− → no DLS
Tiitinen	2003	164	Controls are partly women referred for tubal evaluation	Endometriosis	Case–control study[a]	Only in cases and some controls	Unknown
Alves	2007	55	Cases are partly women referred for tubal evaluation	Antibiotic therapy	Case–control study[a]	Only in cases	Unknown
Guerra-Infante	2003	100	Women referred for tubal evaluation	Surgery <30 days before Concurrent illness	Cohort study (retrospective)	Yes	Unknown
Toye	1993	72	Women referred for subfertility evaluation	Unknown	Case–control study[a]	Not standard	Unknown
Ånestad	1987	103	Cases are women referred for tubal evaluation	Antibiotic therapy	Case–control study[a]	Only in cases	CAT → HSG HSG → DLS
Svenstrup	2007	212	Cases are women referred for tubal evaluation	Unknown	Cohort study (prospective)	Yes	Unknown

Continued

Table I *Continued*

Study	Year	Total women	Inclusion criteria	Exclusion criteria	Study design	DLS performed	Diagnostic strategy
Number of women		6191					

[a]Cases are (partly) women referred for tubal patency testing. Controls were healthy (pregnant) women and were therefore excluded from the analyses.
TT, tubal testing (HSG or DLS); EXP, expectative management for at least 6 months after HSG; HSS, hysterosalpingosonography.
In the last column the diagnostic strategy of the original study was described when reported in the article. Some studies directly performed a DLS, irrespective of the CAT results. Other studies had a different strategy for CAT negative women and CAT positive women. In the same way, management after HSG was in some studies dependent on the obtained results.

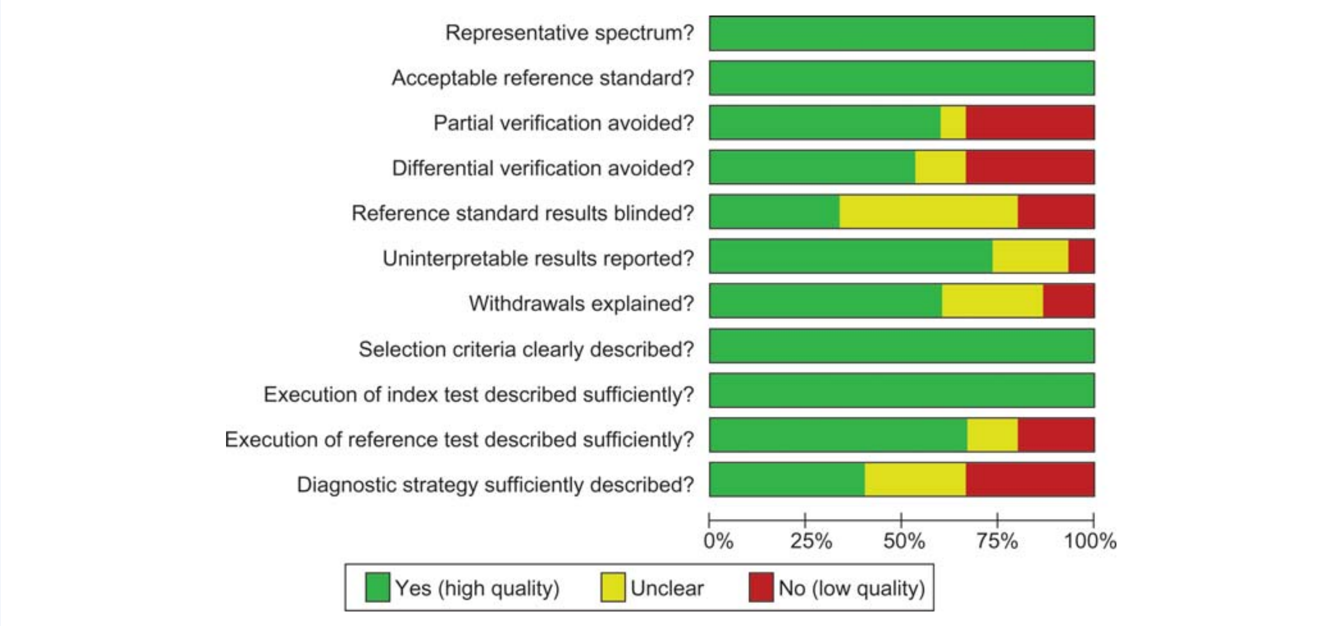

Figure 2 Methodological quality graph: review authors' judgments about methodological quality items presented as percentages of included studies. Overview of methodological quality of reporting of included studies, according to the QUADAS checklist.

Results

Literature search and data collection

In the previous meta-analysis 23 studies had been included (Mol *et al.*, 1997). Our current search for new studies on the subject published after February 1996 resulted in a total of 2642 potential studies on Medline and 1673 potential studies on Embase reporting on tubal pathology, of which 483 were overlapping studies. In total, 283 studies were identified as potentially eligible for inclusion. Of these studies, 164 did not contain diagnostic data on CAT results and/or laparoscopy. Therefore, 119 studies were eligible for inclusion. Checks of cross-references of selected articles identified no other studies. Of the 119 studies, 26 articles were reporting on previously published studies and for 24 studies the authors could not be traced, leaving 69 studies eligible for our IPD meta-analysis.

Of these 69 studies, we contacted the authors and 38 responded. Twenty-four reported their data to be lost, whereas the other 14 authors provided their data (Anestad *et al.*, 1987; Land *et al.*, 1998; Toye *et al.*, 1993; Ng *et al.*, 2001; Mouton, 2002; Veenemans and van der Linden, 2002; Guerra-Infante *et al.*, 2003; Land *et al.*, 2003; Logan *et al.*, 2003; Debattista and Gazzard, 2004; Idahl *et al.*, 2004; den Hartog *et al.*, 2005, 2006; Land and den Hartog, 2006; Tiitinen *et al.*, 2006; Machado *et al.*, 2007; Steures *et al.*, 2007; Svenstrup *et al.*, 2008; van der Steeg *et al.*, 2008). Thirteen studies were single center studies, whereas the largest study was a multi-center study, in which data were collected from 38 centers (Steures *et al.*, 2007; van der Steeg *et al.*, 2008). In the analyses all data from this multicenter study were processed as originating from one study. A flow chart of the included studies is shown in Fig. 1.

Data on 6191 individual women from 14 studies were included in the summary database. For 2872 women, from 12 studies, continuous CAT results were available. For 581 women, from 3 studies, data on two CAT assays per individual woman were available. For the remaining 2738 women there were no continuous CAT results available. These women were excluded from the analyses. Two studies did not contain continuous CAT results and were also excluded from the analysis. Finally, data on 3453 CAT results were available for the analyses. All included women were referred to a fertility clinic after minimal 1 year of unfulfilled child wish. Study characteristics of the included articles are listed in Table I.

Quality assessment

An overview of the methodological quality of the 14 included studies is shown in Fig. 2. All 14 included studies clearly described their patient selection and the execution of CAT. In eight studies the diagnostic strategy followed was not described or unclear. In five studies not all women received the reference test (i.e. DLS); whereas in two studies it was unclear whether all women received laparoscopy and partial verification bias could be avoided.

For 13 studies the comparison of consistency between received data and the published results showed only minimal differences in mean female age and percentage of primary subfertility, which were ignored. One study reported on 295 women instead of 395 women that were included in the database. Therefore, the quality of the received data was considered satisfactory for all included studies.

Statistical analysis

Out of the 3253 women 1686 had CAT assessed with ELISA, 743 with IF and 824 with MIF. In 35% of women with a positive CAT result, and in 50% of women with a negative CAT result, laparoscopy had not been performed. In these 1046 women missing data on laparoscopy were imputed.

The overall prevalences of any tubal pathology and bilateral tubal pathology as defined on DLS were 29% (95% CI 27–30%) and 13% (95% CI 12–15%), respectively. The mean female age, percentages of primary subfertility, median duration of subfertility and prevalences of tubal pathology in the individual studies are shown in Table II. There was a high degree of variability across the studies in duration of subfertility, CAT positivity and prevalence of tubal pathology, which was due to heterogeneity rather than chance as indicated by high I^2 values between 67 and 84%.

Sensitivity and specificity of CAT for the diagnosis of tubal pathology, based on the reported, dichotomized, CAT results per study are shown in Table III. For any tubal pathology sensitivity ranged between 12 and 91% and specificity ranged between 35 and 100%. For bilateral tubal pathology sensitivity was between 31 and 70% and specificity between 52 and 86%.

Both the individual, as well as the pooled AUCs for the diagnosis of any tubal pathology and bilateral tubal pathology are shown in Table IV. The AUCs for the diagnosis of any tubal pathology in the individual studies ranged from 0.52 to 0.79 for the ELISA, from 0.54 to 0.78 for IF and from 0.63 to 0.80 for MIF (see supplementary data). The AUCs for the diagnosis of bilateral tubal pathology ranged from 0.53 to 0.79 for the ELISA, from 0.60 to 0.63 for IF and from 0.77 to 0.78 for MIF.

Table II Overview of prevalences and patient characteristics of the included studies, with assessment of heterogeneity of data across studies (I^2 statistic).

Study	Number of women	Female age mean (5th–95th percentile)	Primary subfertility %	Duration of subfertility median (range)	Prevalence tubal pathology (%)		CAT positivity (%)
					Any tubal pathology	Bilateral tubal pathology	
Bhattacharya	207	30.7 (23–38)	56	2.0 (0.4–10)	30	14	19
Van der Steeg/ Steures	3706	32.4 (25–39)	62	1.6 (0–12)	26	12	29
Ng	110	31.9 (25–37)	75	3.0 (1–12)	27	16	26
Van der Linden	395	31.8 (19–44)	65	2.2 (0.2–13)	27	na	22
Allan	52	33.8 (27–42)	na	na	88	70	14
Mouton	107	34.7 (27–43)	69	na	8	na	30
Land/Den Hartog	662	30.9 (24–37)	72	1.4 (0–9)	18	9	21
Idahl	244	31.1 (24–39)	57	na	43	14	24
Tiitinen	166	32.9 (25–39)	na	na	37	na	32
Alves	55	30.7 (24–38)	na	na	58	44	56
Guerra-Infante	100	29.6 (21–39)	62	na	31	na	na
Toye	72	34.3 (26–42)	na	na	61	na	56
Ånestad	103	30.5 (24–38)	na	na	42	na	76
Svenstrup	212	29.9 (24–37)	na	na	16	na	15
Total	6191				29	13	
Heterogeneity I^2		5%	25%	84%	67%	70%	70%

Table III Overview of sensitivity and specificity of CAT for the individual studies based on the reported cut-off values per study.

Study	CAT assay and cut-off value (manufacturer)	Accuracy of tubal pathology (%)			
		Any TP		Bilateral TP	
		Sens.	Spec.	Sens.	Spec.
Bhattacharya	ELISA 1.1 (Medac)	37	88	50	86
Van der Steeg	ELISA 1.1 (Medac)	57	62	70	61
Steures	IF 1:16 (Biomerieux)	58	62	67	60
Ng	MIF 1:32 (USA)	53	85	67	83
Van der Linden	IF 1:32(Biomerieux)	68	65	na	na
Allan	ELISA na (Labsystems)	12	100	31	76
Mouton	ELISA na (Medac)	69	75	na	na
Land	ELISA 1.1 (Medac)	56	90	60	83
Den Hartog	MIF 1:32 (Biomerieux)	44	78	58	78
Idahl	MIF 1:20 (USA)	42	84	50	77
Tiitinen	ELISA 0.4 (Medac)	42	79	na	na
	MIF na (na)	64	76		
Alves	IF 1:16 (USA)	57	41	70	52
Guerra-Infante	IF 1:8 (USA)	na	na	na	na
Toye	MIF na (na)	73	71	na	na
Ånestad	IF 1:8 (na)	91	35	na	na
Svenstrup	ELISA na (Medac)	23	88	na	na

CAT assay: type of assay used in primary study. In case of the use of several assays in one study, all assays and their accuracies are reported.
na: not applicable.

The empirical pooled ROC analysis showed AUCs for any tubal pathology of 0.57, 0.63 and 0.64 for ELISA, IF and MIF, respectively. For bilateral tubal pathology the AUCs were 0.61, 0.63 and 0.75, for ELISA, IF and MIF, respectively. The random effects logistic regression analysis, in which we accounted for heterogeneity across studies, showed an AUC of 0.64 for ELISA, an AUC of 0.65 for IF and an AUC of 0.75 for MIF. The AUCs of ELISA and MIF, as well as the AUCs of IF and MIF differed significantly (Table IV). Repeating this analysis for bilateral tubal pathology showed AUCs of 0.66 for ELISA, 0.66 for IF and 0.77 for MIF, with significantly different AUCs between ELISA and MIF and between IF and MIF.

The ROC curves of the random effect logistic regression models for the three assays for the diagnosis of any tubal pathology and bilateral tubal pathology are shown in Fig. 3A and B, respectively. The accuracy of MIF was clearly better than the accuracy of ELISA and IF over the full range of sensitivities and specificities.

Discussion

At present, there are several diagnostic strategies for women with suspected tubal pathology. These strategies differ among countries and among fertility centers in the same country and region. Despite clinical research done so far there is still inconsistency between fertility guidelines about the best diagnostic strategy and there is no consensus on which test should be initially used, or on the most effective or cost-effective sequence of tests (British National Collaborating Center for Women's and Children's Health, 2004; Dutch Society of Obstetrics and Gynaecology, 2004). In many fertility centers, CAT is used ahead of more invasive tests such as HSG or laparoscopy, whereas in other centers CAT is not used at all and all women are directly referred for HSG or even laparoscopy.

In this study, we reconsidered the accuracy of three different CAT assays in diagnosing tubal pathology based on an IPD meta-analysis, containing data of 6191 women referred for subfertility in fertility centers all over the world.

The results of this IPD meta-analysis showed that MIF has a moderate ability to discriminate between women with and without tubal pathology. For example, at a sensitivity of 74% MIF had a specificity of 66%, which represents 26% missed cases of tubal pathology and 34% of women incorrectly referred for invasive testing, respectively. Clinically, the optimal cut-off value of CAT depends on the individual context. In some women it will be more important to detect tubal pathology with high certainty, for example, in older women who start IUI treatment. In other women it is more important to rule out tubal pathology in order to avoid unnecessary invasive testing, for example, in couples with a good prognosis for spontaneous pregnancy who delay treatment for a while. The accuracy of MIF is significantly better than the accuracy of ELISA and IF. So far, only a few studies reported on the comparison of several assays and only one compared MIF to ELISA, showing a better performance of the MIF, with a diagnostic odds ratio twice as high as the ELISA test (Land *et al.*, 2003). The analyses in this IPD meta-analysis were based on a large amount of continuous CAT data, not restricted by chosen cut-off values, increasing the consistency and statistical power of this result.

Table IV **Overview of the area under the curves (AUCs) of CAT per assay for the individual studies, as well as for the pooled data.**

Study	Women used in the analyses	AUCs for ELISA		AUCs for IF		AUCs for MIF	
		Any TP	Bilateral TP	Bilateral TP	Any TP	Bilateral TP	Any TP
Bhattacharya	33	0.55	0.53				
Van der Steeg	858	0.52	0.57				
Steures	366			0.56	0.60		
Ng	110					0.71	0.77
Van der Linden	145			0.73	na		
Allan	none						
Mouton	106	0.79	na				
Land	315	0.69	0.79				
Den Hartog	310					0.70	0.78
Idahl	244					0.63	0.78
Tiitinen	162	0.63	na				
	160					0.80	na
Alves	31			0.58	0.63		
Guerra-Infante	100			0.54	na		
Toye	none						
Ånestad	101			0.78	na		
Svenstrup	212	0.53	na				
Pooled area's under the curves							
Empirical pooled (95% CI)		0.57 (0.54–0.60)	0.61 (0.57–0.66)	0.63 (0.59–0.67)	0.63 (0.54–0.71)	0.64 (0.60–0.68)	0.75 (0.68–0.81)
Random effects model* (95% CI)		0.64 (0.62–0.67)	0.66 (0.62–0.70)	0.65 (0.61–0.69)	0.66 (0.59–0.73)	0.75 (0.71–0.78)	0.77 (0.72–0.82)

*P-values between the AUCs as assessed with the random effects model were as followed: Any tubal pathology: Overall: $P < 0.001$; ELISA versus MIF $P < 0.001$, IF versus MIF $P < 0.001$, ELISA versus IF $P = 0.83$. Bilateral tubal pathology: Overall: $P = 0.01$; ELISA versus MIF $P = 0.01$, IF versus MIF $P = 0.01$, ELISA versus IF $P = 0.98$.

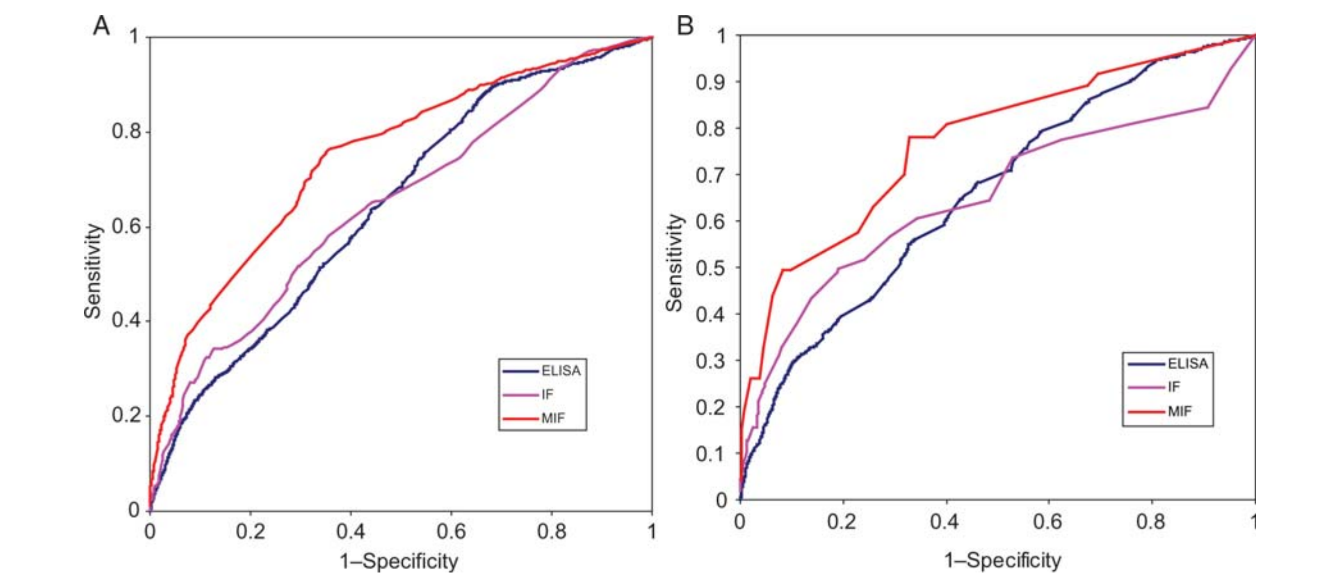

Figure 3 ROC curves for the accuracy of CAT (based on the IPD random effects models) comparison of IPD meta-analysis and conventional meta-analysis for the three assays. (**A**) Accuracy of diagnosing any tubal pathology. The lines replicate the accuracy of CAT based on the IPD meta-analysis, the summary ROC points are based on the conventional meta-analytic approach. Used cut-off values are 1.1 for ELISA, 1:32 for IF and 1:16 and 1:32 for MIF. (**B**) Accuracy of diagnosing bilateral tubal pathology. The lines replicate the accuracy of CAT based on the IPD meta-analysis.

There was a slight difference in accuracy of the three tests for the diagnosis of bilateral tubal pathology and any tubal pathology. The definition of tubal pathology varies in literature. Some studies exclude adhesions or hydrosalpinges, others emphasize the difference between proximal and distal tubal pathology. Our definitions of tubal pathology included occlusion of the fallopian tubes, peritubal adhesions and hydrosalpinges. We did not discriminate between proximal and distal tubal occlusion, since clinical management and pregnancy chances are similar in both types of occlusion (Farhi *et al.*, 2007).

In IPD meta-analyses the substantial number of missing values in the database, both at the study level, as well as on the individual patient level may provide some limitations.

As shown in the flowchart, not all eligible studies could be included in the present meta-analysis. This was due to a lack of information about the authors, a lack of response from the authors or loss of primary data. Combining IPD with aggregate data could overcome the problem of missing studies. Since our primary interest was the use of continuous test results to assess accuracy estimates of CAT, we were not able to use aggregate data, in which continuous test results are not available. To be sure that the selection of women in the included studies was representative, we roughly compared the patient selections in the included studies with the patient selections in the studies that were not available for inclusion. We found no major differences between these patient selections. Also, inclusion of only a subset of eligible studies might lead to biased results, but only if studies that do not participate in the IPD collaboration show test accuracies different to those studies that do participate. Since test accuracy of the included studies varied substantially, we do not expect to find a completely different spectrum in the studies that were not included.

We included a number of case–control studies in the present meta-analysis. In these studies, the cases were women referred for fertility work-up and therefore comparable to women included from cohort studies. The controls were excluded from the meta-analysis. Also, compared with the total number of included women, the included 'cases' accounted for only 6%. We are of opinion that they will not have substantial influence on the estimated accuracy (Rutjes *et al.*, 2005).

Besides missing studies, missing data at the patient level might also be a problem in IPD meta-analyses. Of the 6191 women in the database, we were able to use data of 3453 women in the analyses. A comparison of the characteristics of these women to those that were excluded from the analyses showed no differences.

Missing laparoscopy results is another cause of missing data at the patient level. For 35% of women with a positive CAT result, laparoscopy had not been performed, whereas for 50% of women with a negative CAT results data on laparoscopy were missing. This partial verification happens easily in clinical practice since women with a positive CAT result more often received a laparoscopy compared with women with a negative CAT result. When these subsets of women are not comparable to each other and women with missing laparoscopy results were omitted from the analyses, verification bias would occur (van der Heijden *et al.*, 2006; de Groot *et al.*, 2008). To reduce this partial verification bias we decided to impute the missing laparoscopy results. We decided not to show results from an analysis restricted to women that have both had CAT and laparoscopy, since we would then assume that data were missing completely at random, while we know from clinical practice that CAT results might influence the decision to perform laparoscopy. Also, results from a complete case approach are expected to be more prone to verification bias and could therefore be misleading (Moons *et al.*, 2006; van der Heijden *et al.*, 2006).

At present, there is no consensus on the type of analyses that should be performed in a diagnostic IPD meta-analysis. In a previous study, we described the execution of analyses that we performed in this IPD meta-analysis (Broeze *et al.*, 2010). To be able to express the accuracy of the CAT, we used the area under the ROC curve (AUC). Although sensitivity and specificity might be more applicable measurements for clinicians, these could not be reported without using cut-off levels for the CAT results. To compare the results of the three different assays (MIF, IF and ELISA), without losing any information, we used the continuous test results and therefore reported the AUC. Although there are several CAT assays available worldwide, we only included these three assays, since they were most often used and reported in the original studies and enough data were available for inclusion in this meta-analysis. We used random effects logistic regression models, since pooling data for all women across studies without adjustments would ignore differences across study populations which could have led to biased results. This was also shown in Table IV in which the empirical pooled AUCs for MIF are inconsistent with the data and with the random effects model.

We found relatively low sensitivities of CAT, combined with high specificities. The latter is clinically very important, because this implies that CAT might avoid unnecessary, invasive testing in women without tubal pathology. On the other hand, the estimated sensitivities might be too low to correctly rule out tubal pathology, which implies that women with tubal pathology might remain undiagnosed (Leeflang *et al.*, 2008).

The question whether CAT has clinical relevance as a test in the fertility work-up will also depend on the comparison of its accuracy with that of other diagnostic information, such as medical history. In a previous retrospective study, it was shown that the combination of CAT and medical history taking had a better yield than either of these alone (Coppus *et al.*, 2007). The added value of CAT over the information from medical history, physical examination and other diagnostic tests can be adequately analyzed with IPD and will be assessed in a subsequent study, based on this IPD project. This analysis of the added value of CAT to the pre-test probability of tubal pathology can be restricted to MIF, since this study shows MIF to be the CAT assay with the best discriminative possibility.

In conclusion, our study showed that MIF has a significantly better accuracy than IF and ELISA for the diagnosis of any tubal pathology. The accuracy of MIF showed a moderate ability to discriminate between women with and without tubal pathology. ELISA and IF showed only poor discriminative ability, therefore MIF should be the test of first choice in the initial screen for tubal pathology.

Supplementary data

Supplementary data are available at http://humupd.oxfordjournals.org/.

Authors' roles

The work was performed for the TUBA IPD Sudy group (www.ipd-meta-analysis.com). B.W.M. and F.v.d.V. designed the study.

S.F.P.J.C. and K.A.B. performed the literature search. K.A.B. coordinated this IPD meta-analysis, approached the original authors and created the summary database. M.F.C.A., G.A., S.B., J.A., M.F.G.-I., J.E.D.H, J.A.L., A.I., P.J.Q.v.d.L., J.W.v.d.M., E.H.Y.N., J.W.S., P.S., H.F.S., A.T. and B.T. performed the original data acquisition. K.A.B. did the analysis, under the supervision of B.C.O. and N.V.G. All authors have revised the article and have given final approval of the submitted version.

Funding

This study was financially supported by ZonMW. Grant number 90700201.

References

Anestad G, Lunde O, Moen M, Dalaker K. Infertility and chlamydial infection. *Fertil Steril* 1987;**48**:787–790.

Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics* 1983;**39**:207–215.

British National Collaborating Center for Women's and Children's Health. Fertility: assessment and treatment for people with fertility problems. NICE guideline, 2004.

Broeze KA, Opmeer BC, Bachmann LM, Broekmans FJ, Bossuyt PM, Coppus SF, Johnson NP, Khan KS, ter Riet G, van der Veen F *et al.* Individual patient data meta-analysis of diagnostic and prognostic studies in obstetrics, gynaecology and reproductive medicine. *BMC Med Res Methodol* 2009;**9**:22.

Broeze KA, Opmeer BC, van der Veen F, Bossuyt PM, Bhattacharya S, Mol BW. Individual patient data meta-analysis; a promising approach for evidence synthesis in reporductive medicine. *Human Repod* 2010;**16**:561–567.

Cahill DJ, Wardle PG. Management of infertility. *Br Med J* 2002;**325**:28–32.

Coppus SF, Opmeer BC, Logan S, van der Veen F, Bhattacharya S, Mol BW. The predictive value of medical history taking and Chlamydia IgG ELISA antibody testing (CAT) in the selection of subfertile women for diagnostic laparoscopy: a clinical prediction model approach. *Hum Reprod* 2007;**22**:1353–1358.

Debattista J, Gazzard CM. Interaction of microbiology and pathology in women undergoing investigations for infertility. *Infect Dis Obstet Gynecol* 2004;**12**:135–145.

de Groot JA, Janssen KJ, Zwinderman AH, Moons KG, Reitsma JB. Multiple imputation to correct for partial verification bias revisited. *Stat Med* 2008; **27**:5880–5889.

den Hartog JE, Land JA, Stassen FR, Kessels AG, Bruggeman CA. Serological markers of persistent *C. trachomatis* infections in women with tubal factor subfertility. *Hum Reprod* 2005;**20**:986–990.

den Hartog JE, Morre SA, Land JA. Chlamydia trachomatis-associated tubal factor subfertility: Immunogenetic aspects and serological screening. *Hum Reprod Update* 2006;**12**:719–730.

den Hartog JE, Lardenoije CM, Severens JL, Land JA, Evers JL, Kessels AG. Screening strategies for tubal factor subfertility. *Hum Reprod* 2008;**23**:1840–1848.

DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials* 1986; **7**:177–188.

Dutch Society of Obstetrics and Gynaecology. Dutch guideline: Fertility work-up. 2004.

Evers JL. Female subfertility. *Lancet* 2002;**360**:151–159.

Farhi J, Ben-Haroush A, Lande Y, Fisch B. Role of treatment with ovarian stimulation and intrauterine insemination in women with unilateral tubal occlusion diagnosed by hysterosalpingography. *Fertil Steril* 2007;**88**:396–400.

Guerra-Infante FM, Carballo-Perea R, Zamora-Ruiz A, Lopez-Hurtado M, Flores-Medina S, Contreras GM. Evaluation of an indirect immunofluorescence assay for detecting Chlamydia trachomatis as a method for diagnosing tubal factor infertility in Mexican women. *Int J Fertil Womens Med* 2003;**48**:74–82.

Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982;**143**:29–36.

Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analysis. *Br Med J* 2003;**327**:557–560.

Idahl A, Boman J, Kumlin U, Olofsson JI. Demonstration of Chlamydia trachomatis IgG antibodies in the male partner of the infertile couple is correlated with a reduced likelihood of achieving pregnancy. *Hum Reprod* 2004;**19**:1121–1126.

Land JA, Evers JL, Goossens VJ. How to use Chlamydia antibody testing in subfertility patients. *Hum Reprod* 1998;**13**:1094–1098.

Land JA, Gijsen AP, Kessels AG, Slobbe ME, Bruggeman CA. Performance of five serological chlamydia antibody tests in subfertile women. *Hum Reprod* 2003; **18**:2621–2627.

Land JA, den Hartog JE. Chlamydia antibody testing in subfertile women. *Drugs Today (Barc)* 2006;**42**(Suppl A):35–42.

Leeflang MM, Deeks JJ, Gatsonis C, Bossuyt PM. Systematic reviews of diagnostic test accuracy. *Ann Intern Med* 2008;**149**:889–897.

Logan S, Gazvani R, McKenzie H, Templeton A, Bhattacharya S. Can history, ultrasound, or ELISA chlamydial antibodies, alone or in combination, predict tubal factor infertility in subfertile women? *Hum Reprod* 2003;**18**:2350–2356.

Machado AC, Guimaraes EM, Sakurai E, Fioravante FC, Amaral WN, Alves MF. High titers of Chlamydia trachomatis antibodies in Brazilian women with tubal occlusion or previous ectopic pregnancy. *Infect Dis Obstet Gynecol* 2007;**2007**:24816.

Mol BW, Dijkman B, Wertheim P, Lijmer J, van der Veen F, Bossuyt PM. The accuracy of serum chlamydial antibodies in the diagnosis of tubal pathology: a meta-analysis. *Fertil Steril* 1997;**67**:1031–1037.

Moons KGM, Donders ART, Stijnen T, Harrell FE jr. Using the outcome for imputation of missing predictor values was preferred. *J Clin Epi* 2006; **59**:1092–1101.

Mosher WD, Pratt WF. Fecundity and infertility in the United States: incidence and trends. *Fertil Steril* 1991;**56**:192–193.

Mouton JW. Tubal factor pathology caused by Chlamydia trachomatis: The role of serology. *Int J STD AIDS* 2002;**13**:26–29.

Ng EH, Tang OS, Ho PC. Measurement of serum CA-125 concentrations does not improve the value of Chlamydia trachomatis antibody in predicting tubal pathology at laparoscopy. *Hum Reprod* 2001;**16**:775–779.

Riley RD, Dodd SR, Craig JV, Thompson JR, Williamson PR. Meta-analysis of diagnostic test studies using individual patient data and aggregate data. *Stat Med* 2008;**27**:6111–6136.

Rutjes AWS, Reitsma JB, Vandenbroucke JP, Glas AF, Bossuyt PMM. Case–control and two-gate designs in diagnostic accuracy studies. *Clin Chem* 2005; **51**:1335–1341.

Steures P, van der Steeg JW, Hompes PG, Bossuyt PM, Habbema JD, Eijkemans MJ, Koks CA, Boudrez P, van der Veen F, Mol BW. The additional value of ovarian hyperstimulation in intrauterine insemination for couples with an abnormal postcoital test and a poor prognosis: a randomized clinical trial. *Fertil Steril* 2007;**88**:1618–1624.

Svenstrup HF, Fedder J, Kristoffersen SE, Trolle B, Birkelund S, Christiansen G. Mycoplasma genitalium, Chlamydia trachomatis, and tubal factor infertility—a prospective study. *Fertil Steril* 2008;**90**:513–520.

Taylor A. ABC of subfertility: extent of the problem. *Br Med J* 2003;**327**:434–436.

Tiitinen A, Surcel HM, Halttunen M, Birkelund S, Bloigu A, Christiansen G, Koskela P, Morrison SG, Morrison RP, Paavonen J. Chlamydia trachomatis and chlamydial heat shock protein 60-specific antibody and cell-mediated responses predict tubal factor infertility. *Hum Reprod* 2006;**21**:1533–1538.

Toye B, Laferriere C, Claman P, Jessamine P, Peeling R. Association between antibody to the chlamydial heat-shock protein and tubal infertility. *J Infect Dis* 1993;**168**:1236–1240.

Tuuminen T, Palomaki P, Paavonen J. The use of serologic tests for the diagnosis of chlamydial infections. *J Microbiol Methods* 2000;**42**:265–279.

van der Heijden GJMG, Donders ART, Stijnen T, Moons KG. Imputation of missing values is superior to complete case analysis ans the missing-indicator method in multivariable diagnostic research: A clinical example. *J Clin Epi* 2006; **59**:1102–1109.

van der Steeg JW, Steures P, Eijkemans MJ, Habbema JD, Hompes PG, Michgelsen HW, van der Heijden PF, Bossuyt PM, van der Veen F, Mol BW. Predictive value of pregnancy history in subfertile couples: results from a nationwide cohort study in the Netherlands. *Fertil Steril* 2008;**90**:521–527.

Veenemans LM, van der Linden PJ. The value of Chlamydia trachomatis antibody testing in predicting tubal factor infertility. *Hum Reprod* 2002;**17**:695–698.

Wang SP, Grayston JT. Immunologic relationship between genital TRIC, lymphogranuloma venereum, and related organisms in a new microtiter indirect immunofluorescence test. *Am J Ophthalmol* 1970;**70**:367–374.

Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Med Res Methodol* 2003;**3**:25.