Journal of Clinical Epidemiology 68 (2015) 290–298

Using individual patient data to adjust for indirectness did not successfully remove the bias in this case of comparative test accuracy

Junfeng Wang[a], Patrick Bossuyt[a], Ronald Geskus[a], Aeilko Zwinderman[a], Madeleine Dolleman[b], Simone Broer[b], Frank Broekmans[b], Ben Willem Mol[c], Mariska Leeflang[a,*], on behalf of the IMPORT Study Group

[a]*Department of Clinical Epidemiology, Biostatistics & Bioinformatics, Academic Medical Center, PO Box 22700, 1100 DE Amsterdam, The Netherlands*
[b]*Department of Reproductive Medicine, University Medical Center Utrecht, PO Box 85500, 3508 GA Utrecht, The Netherlands*
[c]*Department of Obstetrics and Gynaecology, Women's and Children's Hospital, University of Adelaide, 72 King William Road, SA 5006 Adelaide, Australia*

Accepted 17 October 2014; Published online 2 December 2014

Abstract

Objectives: In comparative systematic reviews of diagnostic accuracy, inconsistencies between direct and indirect comparisons may lead to bias. We investigated whether using individual patient data (IPD) can adjust for this form of bias.

Study Design and Setting: We included IPD of 3 ovarian reserve tests from 32 studies. Inconsistency was defined as a statistically significant difference in relative accuracy or different comparative results between the direct and indirect evidence. We adjusted for the effect of threshold and reference standard, as well as for patient-specific variables.

Results: Anti-Müllerian hormone (AMH) and follicle stimulation hormone (FSH) differed significantly in sensitivity (−0.1563, $P = 0.04$). AMH and antral follicle count (AFC) differed significantly in sensitivity (0.1465, $P < 0.01$). AMH and AFC differed significantly in specificity (−0.0607, $P = 0.02$). The area under the curve (AUC) differed significantly between AFC and FSH (0.0948, $P < 0.01$) in the direct comparison but not (0.0678, $P = 0.09$) in the indirect comparison. The AUCs of AFC and AMH differed significantly (−0.0830, $P < 0.01$) in the indirect comparison but not (−0.0176, $P = 0.29$) in the direct comparison. These differences remained after adjusting for indirectness.

Conclusion: Estimates of comparative accuracy obtained through indirect comparisons are not always consistent with those obtained through direct comparisons. Using IPD to adjust for indirectness did not successfully remove the bias in this case study.

Keywords: Diagnostic test accuracy; Comparative meta-analysis; Individual patient data; Sensitivity and specificity; Receiver operating characteristic; Generalized estimating equation

1. Introduction

Studies of test accuracy evaluate how well a test is able to identify patients with the target condition, or target event, by comparing test results against the reference standard. Systematic reviews of test accuracy studies try to obtain more precise summary estimates of the accuracy and to explore sources of variability in accuracy. Some reviews target not just one medical test but two or more and evaluate whether the accuracy of one test is better than that of another one. In such comparative systematic reviews, one can include direct and indirect test comparisons. Direct comparisons, also known as head-to-head comparisons, evaluate two or more tests in the same study, preferably in the same patients. Indirect comparisons refer to data from separate studies: one test is evaluated in a series of studies, whereas the second test is evaluated in different studies and different patients.

For various reasons, for example, different test settings, different patients, indirect comparisons are more prone to bias than direct comparisons, and one may be tempted to restrict comparative reviews to direct comparisons [1]. On the other hand, excluding indirect comparisons in systematic reviews may lead to a loss in precision in the summary estimates and fewer data to explore heterogeneity.

Conflict of interest: None.
Funding: This project was supported by The Netherlands Organization for Scientific Research (project 916.10.034).
* Corresponding author. Tel.: +31-205666934; fax: +31-206912683.
E-mail address: m.m.leeflang@amc.uva.nl (M. Leeflang).

http://dx.doi.org/10.1016/j.jclinepi.2014.10.005
0895-4356/

What is new?

Key findings

- Comparative results of test accuracy obtained through indirect comparisons are not always consistent with those obtained through direct comparisons. Even with individual patient data (IPD), there is no generally applicable way to make results of indirect comparisons more comparable to results of direct comparisons.

What this add to what was known?

- All previous studies on indirectness in comparative systematic review were based on study-level data. This is the first time IPD is used to investigate and adjust for indirectness.

What is the implication and what should change now?

- It is difficult to get unbiased estimates from indirect comparisons, even if with adjustment on IPD level. A comparative study design in diagnostic test accuracy studies can make the comparisons more reliable.

Inconsistency in the treatment effects between direct and indirect comparisons has previously been observed in systematic reviews of competing interventions [2]. This finding also applies to systematic reviews of diagnostic test accuracy. Takwoingi et al. [3] compared results from direct and indirect comparisons of diagnostic tests in 36 reviews and found that indirect comparisons do give different results than direct comparisons and the direction of the bias cannot be predicted.

Ways to correct for indirectness were investigated by several researchers. Leeflang et al. analyzed 17 comparisons between assays for D-dimer testing and found a significant effect of indirectness in five of them. To make results from indirect comparisons in correspondence with results from direct comparisons, they used a bivariate random-effects meta-regression model with assay-type and directness as covariates and included study features to correct for the effect of indirectness on sensitivity or specificity. The results in the study by Leeflang et al. [4] showed that adjusting for study features did not have much effect on removing the indirectness. So, it is still doubtful whether and how direct and indirect comparisons in systematic reviews and meta-analysis of test accuracy studies can be combined successfully, that is, without introducing bias.

All previous studies were based on aggregated data at study level, which vary with the threshold for test positivity, the clinical reference standard, and the target population. This information can often be obtained from primary studies. An advanced approach to summarizing the evidence from primary studies is to acquire the original data from included studies and to perform statistical analyses at the individual patient data (IPD) level. IPD meta-analysis offers the possibility of performing additional types of analyses, such as reconciling thresholds and reference standards from primary studies to the same value, adjusting for baseline differences in study-level as well as patient-level characteristics, and using continuous results instead of dichotomized cutoff values [5].

The objective of this case study was to investigate whether using IPD from primary studies can overcome the limitations in analyses based on study-level data. We explored how we can adjust for indirectness with IPD meta-analysis and developed and evaluated methods for adjusting the indirect comparisons, so that the results from such comparisons are more consistent with those from direct comparisons.

2. Data

2.1. Data acquisition

This IPD case study was facilitated by the EXPORT data set used in the ‘‘Excessive Response Prediction using Ovarian Reserve Tests’’ project, a collaborative IPD meta-analysis comparing the accuracy of anti-Müllerian hormone (AMH), antral follicle count (AFC), and follicle stimulation hormone (FSH) in predicting poor ovarian response in in vitro fertilization (IVF) [6]. The data set contained 34 databases including 6,852 women undergoing IVF.

These ovarian reserve tests (ORT) were initially suggested to have a good predictive value for pregnancy, but recent studies showed that these tests are more effective in predicting the ovarian response [7]. AMH, AFC, and FSH are three most widely used ORTs frequently used before IVF treatment to predict poor response to ovarian stimulation [8].

Patient characteristics, such as age, body mass index (BMI), or duration of subfertility, not only have a strong predictive power for ovarian response but also influence the inherent discriminatory accuracy of the ORTs [6]. These variables can help in finding out whether the difference in baseline characteristics is the source of bias in indirect comparisons and provide us the probability to adjust for indirectness by including covariates.

Comparisons were limited to pairs of tests, which are the simplest and most common cases of test comparison. So from the data set, we can generate three pairwise comparisons between two tests: AMH vs. FSH, AMH vs. AFC, FSH vs. AFC, which could make best use of the IPD data set and provide more evidence to evaluate the usefulness of the adjustments. In each pairwise comparison, a direct comparison was defined as a study in which patients had

taken both tests; an indirect comparison was defined as one in which patients had undergone only one of the two tests.

2.2. Dichotomous tests and continuous tests

Some diagnostic tests have only two possible results, classified as positive and negative, and such tests are termed dichotomous tests. Other tests with continuous results are termed continuous tests and may provide useful clinical information over a wide range of values. Diagnostic accuracy of dichotomous tests can be expressed in sensitivity and specificity, likelihood ratios, and diagnostic odds ratio, whereas the discriminatory power of continuous tests is usually measured with the area under a receiver operator characteristic (ROC) curve [9].

Many test results are continuous in nature but classified as positive and negative; thus, in most of the diagnostic test accuracy (DTA) studies, data are generally reported in a dichotomous way, that is, in 2 × 2 tables. This common and simple way of reporting provides reduced information for meta-analyses and neglects the potential diagnostic information contained in continuous test results. Different formats of reporting in primary studies will lead to different statistical methods implemented in data analysis. So, the three ORTs are treated as both dichotomous tests and continuous tests, which will be discussed separately in following sections.

3. Methods

There are two main sources of bias in indirect comparisons: heterogeneity between studies, which may be from different reference standards or thresholds in primary studies, and differences in baseline characteristics, which may lead to confounding and effect modification. We propose two corresponding types of adjustments with IPD. One focuses on the comparability of test results from different studies (type I) and the second on covariate effect (type II). The two adjustments are on different layers: only test results (index tests, reference standard, or both) are needed for type I adjustment, which are the essential information from the primary studies; more information, for example, patient characteristics, are needed for type II adjustment. These two types of adjustments could be performed individually or together. When there are no sufficient patient-level data containing patient characteristics, thus type II adjustment is not feasible, we can use only type I adjustment and vice versa. But it is highly recommended to perform type I adjustment all the time as the first step when it is possible because adjustment on test results can influence the estimate of test accuracy directly. So, in the analyses of this case study, type I adjustment was implemented in analysis 1 and type I + type II adjustments were implemented in analysis 2.

3.1. Type I adjustment: adjustment of reference standard and test results

In meta-analysis of DTA studies, the included primary studies may use different reference standards or use the same reference standard but with different cutoff values to define diseased and nondiseased patients. This difference may lead to heterogeneity in test accuracy. IPD provides the opportunity to redefine the disease status of all patients if individual-level information about the reference standard was reported in the data set. So by adjusting reference standard, we can make sure that test accuracies in different primary studies are measured against the same reference standard and the same cutoff point.

Besides the reference standard, the definition of the positivity of index tests may also vary among studies and the differences in sensitivity and specificity between studies may result from the use of different threshold levels [10]. To make the pooling of data from primary studies more comparable, for each index test, a single cutoff value should be defined and applied to all the patients in all studies. The general cutoff point of index test can be obtained by maximizing overall accuracy or minimizing the total cost of misclassification, and this value should be reasonable and in the range of cutoffs reported in the primary studies.

For continuous tests, there are no cutoffs, but test results can differ between primary studies both in controls and cases. Janes and Pepe [11] proposed a model to correct for the heterogeneity, by standardizing the test results for differences in the test result levels in controls between studies. They use the distribution of continuous test results in the control population as a reference distribution and calculate ‘‘percentile value’’ by standardizing the test results in the case population. Percentile values do not have measurement units and take values between 0 and 1; thus, systematic differences in index test results can be removed by using percentile value instead of original value. In our analysis, percentile values will be used in ROC analysis and calculation of AUC.

3.2. Type II adjustment: adjustment of covariate effect

We can alternatively perform covariate adjustment with patient characteristics that may influence the test accuracy. We can call this kind of adjustments as type II adjustments.

When there are covariates associated with both test results and disease status, test accuracy may be overestimated or underestimated if these confounders are not considered in the design of diagnostic accuracy studies [12]. In direct comparisons, all the tests are evaluated based on the same patient population, so the comparison is less affected by these covariates, but in indirect comparisons, tests are evaluated in different patient groups, which may have different level of confounders, for example, age distribution. Thus, comparative results may be distorted in indirect comparisons. Regression models can be used for exploration of

factors that influence diagnostic test accuracy (sensitivity and specificity) by including covariates [13]. However, maybe due to the lack of a universal method for adjusting for covariates, controlling for confounding has been rarely used by clinical investigators in the context of diagnostic studies [12].

We first consider dichotomous tests. In systematic reviews of dichotomous tests, meta-analysis of sensitivity and specificity values is preferred over meta-analysis of likelihood ratios [14], so we compare dichotomous tests by their sensitivities and specificities. Pooling of sensitivity and specificity separately is not recommended because the paired nature of sensitivities and specificities from individual studies is ignored. A robust and commonly used approach is the construction of a summary receiver operating characteristic curve, which considers the underlying relationship between sensitivity and specificity. However, in this IPD meta-analysis, we combined data but not estimates from individual studies, and sensitivity and specificity were defined on a per-observation basis, so they can be analyzed separately.

When comparing two tests, some patients provide data for only one test; test results from a single patient who took two tests are correlated. A marginal regression model framework proposed by Leisenring et al. [13] allows comparing diagnostic tests with unbalanced data, which contain both paired data (direct comparison) and unpaired data (indirect comparison), and generalizes McNemar's test for paired binary data.

In this study, we have patient characteristics such as age, BMI level, and duration of subfertility as covariates for type II adjustment. The parameter estimation was implemented with generalized estimating equations (GEE) with exchangeable correlation structure. By comparing parameters from a model including these covariates with parameters of the basic model without covariates, we can investigate whether covariate adjustment is a way to correct for indirectness.

For continuous tests, AUC was used as a measure of test accuracy. Janes and Pepe [11] showed that when confounding was present, the overall ROC curve and AUC substantially differed from stratum-specific ROC curve and AUC. Thus, they suggested that methods for covariate adjustment are needed in ROC analysis. In this study, adjusting for covariate effect is implemented with covariate-adjusted ROC (AROC) curve [15,16]. With this model, covariates that may influence the test accuracy could be statistically adjusted in the ROC analysis.

3.3. Comparing diagnostic test accuracy in direct and indirect comparisons

For dichotomous tests, we include binary variable Z that indicates indirectness in comparison and the interaction term with test type, then the hypothesis $H_0 : \beta_D^3 = 0$ ($H_0 : \beta_{\overline{D}}^3 = 0$), where β^3 is the parameter of the interaction term in formula 1 (or formula 2) (see Appendix A at www.jclinepi.com), is equivalent to a statement that there is no difference between direct and indirect comparisons of test sensitivity (1 − specificity).

For continuous tests, in each pair of comparisons, AUCs of each test were estimated with empirical (nonparametric) method in direct comparison and indirect comparison separately and compared with DeLong's test for two correlated ROC curves in paired data (direct comparisons) and its extension for unpaired ROC curves in indirect comparisons [17]. In type II adjustments, AROC was used as the measure of accuracy instead of AUC. Then, we can see whether there are inconsistencies between comparative results from direct and indirect comparisons.

4. Results

4.1. Data set

The final data set only included women who provided information about ovarian response, in terms of number of oocytes, who had taken at least one of the three ORTs. As a result, 4,762 women from 32 databases were suitable for the analysis of tests comparison, in which 1,001 (21.0%) women had a poor response. Table 1 lists the number of patients in each pairwise comparison separated by the type of comparisons.

4.2. Reference standard

Different cutoff points were used in primary studies to define "poor response." Because in our IPD data set, every study reported the exact number of oocytes for each individual patient, we defined poor response for all individual patients according to a single and commonly used definition: the yield of four or less oocytes at follicle [18].

4.3. Dichotomous tests

In this study, we defined the single cutoff value for each ORT in all studies by maximizing the Youden's index [19] (sensitivity + specificity − 1) based on the data set and consulting with recently published systematic reviews of each ORT [20–22]. We know that using Youden's index is debatable, but this way we have an objective and uniform method for selecting a cutoff. AMH test results less than 1.28 ng/mL were defined as positive; AFC number less than

Table 1. Number of patients in each pairwise comparison

	FSH vs. AFC		AMH vs. AFC		FSH vs. AMH	
Comparisons	**FSH**	**AFC**	**AMH**	**AFC**	**FSH**	**AMH**
Direct comparison	2,248		1,024		1,747	
Indirect comparison	2,108	252	867	1,476	2,609	144

Abbreviations: FSH, follicle stimulation hormone; AFC, antral follicle count; AMH, anti-Müllerian hormone.

or equal to 8 were defined as positive; FSH test results larger than or equal to 7.72 IU/L were defined as positive. These values are all in the range of cutoffs reported in the systematic reviews of these ORTs [20–22].

Comparisons between the tests sensitivities are based on the marginal regression model in formula 1 (see Appendix A at www.jclinepi.com), where $Z = 1$ for indirect comparisons and $Z = 0$ for direct comparisons, XZ is the interaction term of test type and indirectness. A similar model (formula 2, Appendix A at www.jclinepi.com) is implemented for 1 − specificity. Analysis 1 represents comparative results from data adjusted for reference standard and threshold effect (type I adjustment), and analysis 2 represents comparative results after adjusting for covariate effect in addition to analysis 1 (type I + type II adjustments). Both results are shown in Table 2.

Table 2 lists the parameter estimates and 95% confidence intervals (in parentheses). For sensitivity, the differences between AMH vs. FSH (−0.1563, $P = 0.04$) and AMH vs. AFC (0.1465, $P < 0.01$) in direct and indirect comparisons are significant. For specificities, the difference between AMH vs. AFC (−0.0607, $P = 0.02$) in direct and indirect comparisons is significant. Thus, after applying the same reference standard and thresholds to all primary studies, inconsistency between direct and indirect comparisons is still observed. This means that in this case, with type I adjustment only, we cannot always remove the bias of indirectness.

We further adjusted the models by including the following covariates in the regression models for sensitivity (1 − specificity): age and accordingly the interaction term of age and test type (analysis 2). After type II adjustment, the parameter β^3, which indicates the differences in sensitivities or specificities, is still significant. The results we got from analysis 2 showed that the inclusion of patient characteristic had no influence on those comparisons, and with type II adjustment, we cannot remove the bias of indirectness either.

4.4. Continuous tests

The ROC curves and AUCs after adjusting for reference standard and standardizing the test results (type I adjustments) are shown in Fig. 1 and Table 3. In the first pair, both direct comparisons and indirect comparisons gave the same conclusion: AMH had a better performance than FSH, but the discriminatory power of both tests in direct comparisons was higher than in indirect comparisons. In the second pair, the difference in AUCs between AFC and FSH (0.0948, $P < 0.01$) is significant in direct comparisons but not significant (0.0678, $P = 0.09$) in indirect comparisons. It was observed that AFC performed much better when directly compared with FSH. In the third pair, the difference between AFC and AMH is significant (−0.0830, $P < 0.01$) in indirect comparison but not significant (−0.0176, $P = 0.29$) in direct comparison. This is because of the increase from 0.78 to 0.83 in the AUC of AMH and the drop from 0.76 to 0.75 in the AUC of AFC in the indirect comparison. The inconsistencies were observed after type I adjustments.

Table 2. Regression coefficients for analyses of sensitivity and 1 − specificity

	Sensitivity		1 − Specificity	
Comparisons	**Analysis 1**	**Analysis 2**	**Analysis 1**	**Analysis 2**
AMH vs. FSH				
β_0 Intercept	0.4277 (0.2388, 0.6166)	−1.2196 (−2.3203, −0.1188)	−1.0660 (−1.1907, −0.9412)	−3.7703 (−4.3701, −3.1705)
β_1 Test type	0.7898 (0.5422, 1.0374)	−3.3739 (−5.2030, −1.5448)	−0.0827 (−0.2447, 0.0794)	−3.5339 (−4.7672, −2.3007)
β_2 Indirectness	−0.1661 (−0.4321, 0.0999)	−0.2127 (−0.4832, 0.0579)	0.2819 (0.1276, 0.4363)	0.1465 (−0.0134, 0.3064)
β_3 Test type × indirectness	**−0.7491 (−1.4489, −0.0492)**	**−0.9264 (−1.6251, −0.2277)**	−0.0362 (−0.4993, 0.4270)	−0.2241 (−0.6698, 0.2217)
β_4 Age		0.0453 (0.0151, 0.0754)		0.0811 (0.0636, 0.0986)
β_5 Test type × age		0.1177 (0.0656, 0.1699)		0.1006 (0.0656, 0.1357)
AMH vs. AFC				
β_0 Intercept	0.8755 (0.5965, 1.1545)	−4.7563 (−6.3721, −3.1404)	−1.1363 (−1.2993, −0.9733)	−5.4416 (−6.2868, −4.5965)
β_1 Test type	−0.0202 (−0.3079, 0.2676)	0.1319 (−1.8282, 2.0919)	−0.1891 (−0.3669, −0.0112)	−1.8664 (−3.0744, −0.6583)
β_2 Indirectness	−0.1940 (−0.5751, 0.1870)	−0.3086 (−0.7147, 0.0975)	0.0937 (−0.1135, 0.3009)	−0.0231 (−0.2376, 0.1914)
β_3 Test type × indirectness	**0.7650 (0.2675, 1.2625)**	**0.8967 (0.3619, 1.4316)**	**0.3279 (0.0477, 0.6080)**	**0.3419 (0.0445, 0.6393)**
β_4 Age		0.1573 (0.1125, 0.2022)		0.1289 (0.1047, 0.1531)
β_5 Test type × age		−0.0044 (−0.0604, 0.0517)		0.0485 (0.0141, 0.0828)
FSH vs. AFC				
β_0 Intercept	0.7563 (0.5540, 0.9587)	−5.1348 (−6.8609, −3.4088)	−1.1552 (−1.2630, −1.0474)	−5.3853 (−6.2536, −4.5171)
β_1 Test type	−0.5519 (−0.7970, −0.3069)	3.5902 (1.7434, 5.4370)	−0.0182 (−0.1589, 0.1224)	1.5623 (0.5567, 2.5678)
β_2 Indirectness	0.1375 (−0.4473, 0.7223)	0.4413 (−0.1497, 1.0323)	0.7057 (0.3948, 1.0166)	0.7387 (0.4166, 1.0608)
β_3 Test type × indirectness	0.1360 (−0.4936, 0.7651)	−0.1561 (−0.7910, 0.4788)	−0.1354 (−0.4746, 0.2038)	−0.2599 (−0.6127, 0.0928)
β_4 Age		0.1617 (0.1147, 0.2087)		0.1246 (0.0998, 0.1493)
β_5 Test type × age		−0.1144 (−0.1648, −0.0640)		−0.0460 (−0.0748, −0.0173)

Abbreviations: AMH, anti-Müllerian hormone; FSH, follicle stimulation hormone; AFC, antral follicle count.

Numbers in parentheses are 95% confidence intervals of parameter estimates; significant differences between direct and indirect comparisons are in bold.

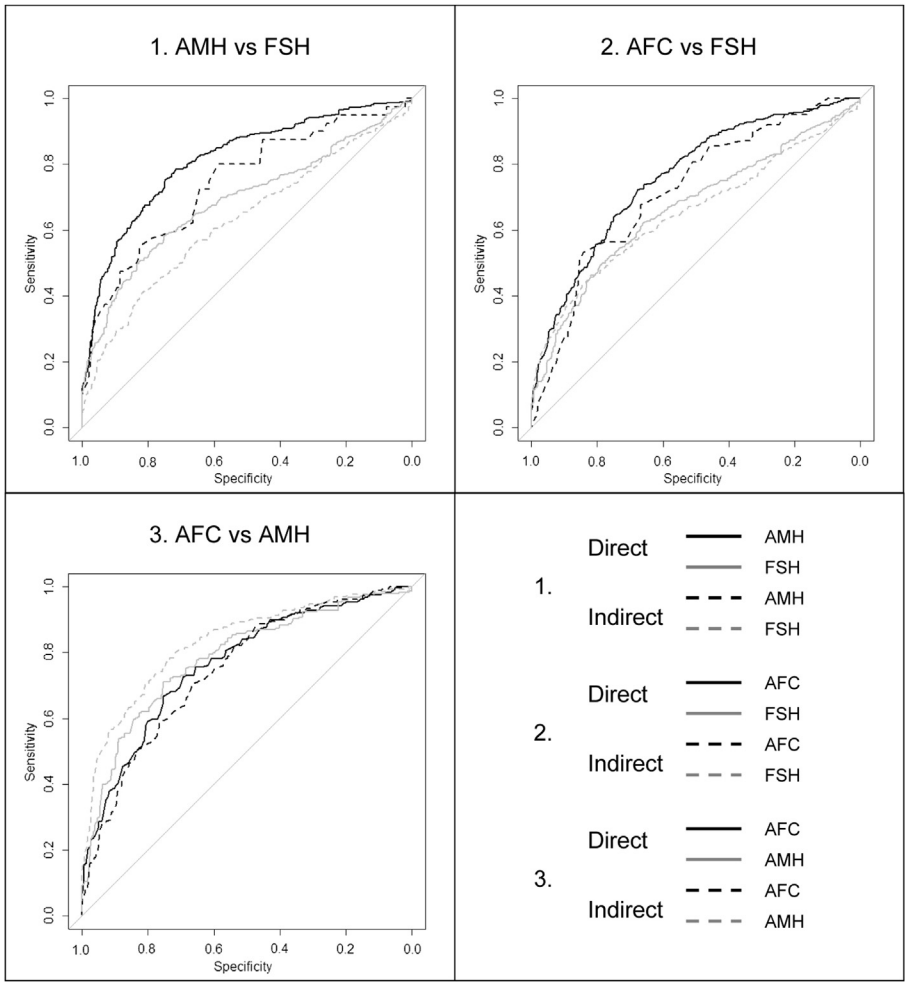

Fig. 1. ROC curves of pairwise comparisons in direct and indirect comparisons after adjustment for heterogeneity.

Fig. 2 and Table 4 show the ROC curves and AUCs after adjusting for covariates (type II adjustments) in addition to type I adjustments. The inconsistencies between direct and indirect comparisons still existed after type II adjustments, in which we tried to adjust for indirectness by considering covariate effect.

5. Discussion

In this IPD case study, we proposed two types of adjustments to correct for the effect of indirectness in comparative systematic reviews of DTA studies. Type I adjustments were focused on threshold effect and reference standard issues, whereas type II adjustments additionally focused on patient characteristics. These adjustments were not successful in removing the bias from indirectness in the present case study: differences between direct and indirect comparisons persisted, even after applying these adjustments.

Analyses were performed with both dichotomous tests and continuous tests. Because most of the DTA studies report dichotomized test results, ORT results were firstly treated as dichotomous tests and compared by their sensitivities and specificities using the generalized McNemar's score test. The bias was defined as the difference in relative accuracy between direct and indirect comparisons and tested by the significance of the parameter in a GEE model. It is a very intuitive and powerful way to detect the difference between direct and indirect comparisons. In the three pairs of comparisons, two pairs showed a significant effect of indirectness on sensitivity and/or specificity both after type I and type II adjustments.

If we keep the continuous nature of ORT data and compare ROC curves generated from direct and indirect comparisons, we also observed these inconsistencies. Although the difference in the second pair (AFC vs. FSH) may attribute to the stronger power of the statistical test used in paired data, because small difference in AUCs can be significant if they are strongly correlated, the difference in the third pair (AFC vs. AMH) confirmed our finding. Neither type I nor type II adjustments can remove the bias of indirectness successfully.

Table 3. Comparisons of AUCs in each pairwise comparison

AMH vs. FSH	**AMH**	**FSH**	**Difference**	***P*-value**
Direct comparison	0.8135 (0.79, 0.84)	0.6888 (0.66, 0.72)	0.1247	**<0.001**
Indirect comparison	0.7434 (0.65, 0.84)	0.6280 (0.60, 0.66)	0.1154	**0.0246**
AFC vs. FSH	**AFC**	**FSH**	**Difference**	***P*-value**
Direct comparison	0.7606 (0.74, 0.79)	0.6658 (0.63, 0.70)	0.0948	**<0.001**
Indirect comparison	0.7202 (0.65, 0.79)	0.6524 (0.62, 0.68)	0.0678	0.0925
AFC vs. AMH	**AFC**	**AMH**	**Difference**	***P*-value**
Direct comparison	0.7648 (0.73, 0.80)	0.7824 (0.75, 0.82)	−0.0176	0.2905
Indirect comparison	0.7474 (0.72, 0.78)	0.8304 (0.80, 0.86)	−0.0830	**0.0003**

Abbreviations: AUC, area under the curve; AMH, anti-Müllerian hormone; FSH, follicle stimulation hormone; AFC, antral follicle count.
Bold indicates significance level = 0.05

Adjustment for indirect comparison in DTA meta-analysis is not as successful as in conventional meta-analysis of interventions. For intervention meta-analysis, Song et al. [23] found that indirect comparison with adjustment may be less biased than direct comparison. This finding may not hold for indirect comparison in DTA meta-analysis, given the different features of DTA and intervention meta-analyses. Sometimes, people may be too optimistic and overestimate the power of adjustment, when we meet the problem of bias from indirect comparison.

In this study, comparative results from direct comparisons are assumed as the gold standard and not biased. But whether this assumption is valid cannot be tested.

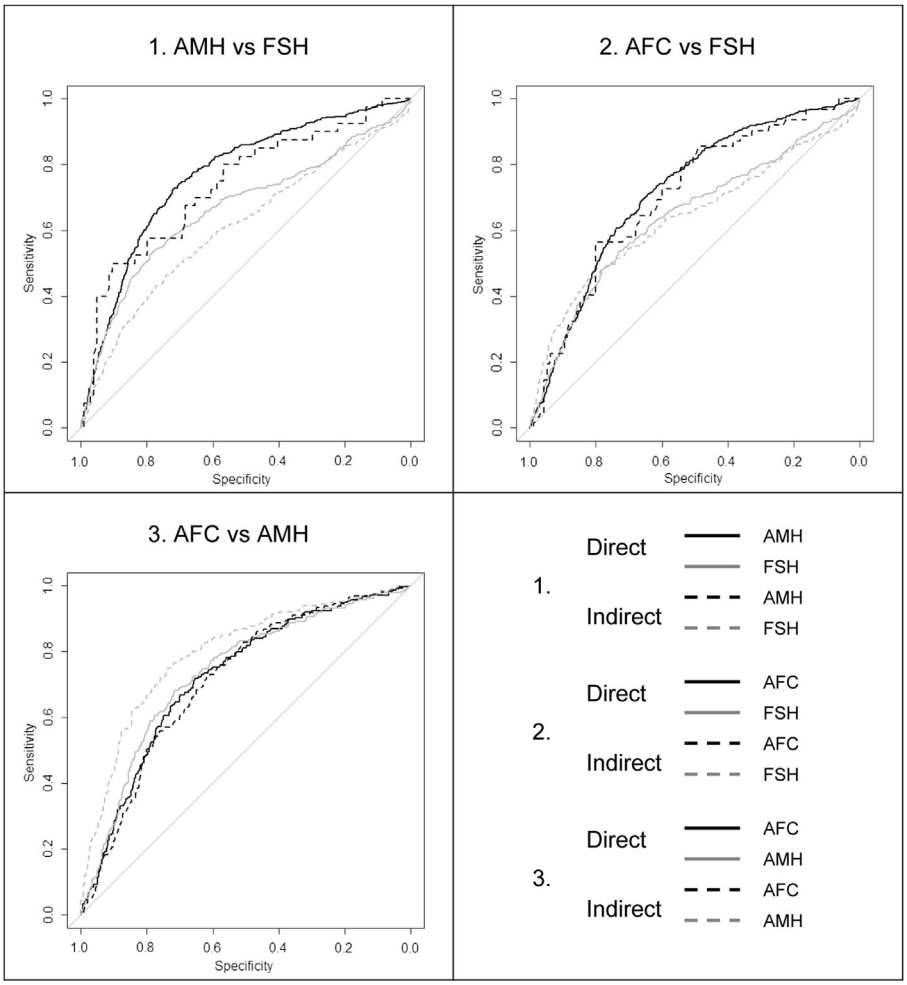

Fig. 2. ROC curves of pairwise comparisons in direct and indirect comparisons after covariate adjustment of age.

Table 4. Comparisons of AUCs in each pairwise comparison after covariate adjustment of age

AMH vs. FSH	AMH	FSH	Difference	*P*-value
Direct comparison	0.7669 (0.74, 0.79)	0.6661 (0.63, 0.70)	0.1008	<**0.001**
Indirect comparison	0.7374 (0.64, 0.83)	0.6095 (0.58, 0.64)	0.1279	**0.0135**
AFC vs. FSH	**AFC**	**FSH**	**Difference**	***P*-value**
Direct comparison	0.7193 (0.69, 0.74)	0.6424 (0.61, 0.67)	0.0769	<**0.001**
Indirect comparison	0.7051 (0.63, 0.78)	0.6375 (0.60, 0.67)	0.0676	0.0985
AFC vs. AMH	**AFC**	**AMH**	**Difference**	***P*-value**
Direct comparison	0.7207 (0.68, 0.76)	0.7332 (0.70, 0.77)	−0.0125	0.4797
Indirect comparison	0.7127 (0.68, 0.75)	0.7949 (0.76, 0.83)	−0.0822	**0.0006**

Abbreviations: AUC, area under the curve; AMH, anti-Müllerian hormone; FSH, follicle stimulation hormone; AFC, antral follicle count.
Bold indicates significance level = 0.05.

There are also other limitations of our research. First, for certain types of diagnostic tests, such as computed tomography and magnetic resonance imaging, which do not have their thresholds, type I adjustments are not always feasible. Second, in type II adjustments, the covariate effect was considered in a linear fashion. However, the mechanism how covariates affect test accuracy may be more complex. Third, although we have age, BMI, and duration of subfertility in this data set, there were too many missing values in BMI and duration of subfertility. Thus, we could only use age in type II adjustments in the analysis. It is also possible that there is another important confounder, which is not in our data set or even not known. Fourth, methods for doing DTA IPD meta-analysis vary among systematic reviews because of the different aims of the reviews and data available. Because test results in the same patients are correlated and comparison between tests is the main interest, in this study we used the GEE approach and paired ROC comparison, which are focused on paired data. It is also possible to compare sensitivities on certain value of specificity or vise versa between two tests. This information can be observed from the ROC curves. Last but not least, this study is only a case study. Because IPD is seldom available, especially for which comparing two or more tests and contains both direct and indirect comparisons, we have only one data set to perform the analysis. We believe that further research needs to be done to investigate whether the bias can be removed if we can adjust for all relevant covariates, which was not the case in the current empirical study.

Unlike effect of intervention studies, in which there is always a control group or a competitor intervention, DTA studies can only evaluate one single test against reference standards. Although researchers always have some comparisons in mind, either comparing the new one with the old one or the one from the neighbors, a competitor test is not a must in study design [24]. In this study, we found it is difficult to get unbiased estimates from indirect comparisons, even if with adjustment on IPD level. The differences found between direct and indirect comparisons may lead to a different decision in clinical practice. For example, the difference found in logit sensitivity between AMH and FSH (Table 2) will mean that in the indirect comparisons, both tests have more or less the same sensitivity, around 57%. In the direct comparisons, AMH will have a higher sensitivity (77%) than FSH (60%). That may mean that if one has to decide what test to rely on more, the indirect comparisons may lead to a decision that it does not matter, whereas the direct comparisons may lead to the decision of AMH being the test to rely on. The same can be seen in the comparison between FSH and AFC but this time the other way around. Although the difference is not statistically significant, the results from the direct comparison (both tests similar accuracy) may lead to a different decision than the results from the indirect comparisons (AFC higher sensitivity than FSH). Thus, to get a more reliable comparison result and better evidence to support the decision, a comparative study design in DTA studies is needed. Systematic reviews will also benefit from better design of primary studies. Comparative studies should be encouraged in DTA studies.

6. Conclusion

Comparative results of test accuracy obtained through indirect comparisons are not always consistent with those obtained through direct comparisons. Study-level covariates were considered for adjusting the bias of indirectness, but the adjustment did not successfully solve the problem. Systematic reviews with IPD are considered as the gold standard, but even with IPD, type I and type II adjustments still cannot remove the bias of indirectness successfully. There is no generally applicable way to make results of indirect comparisons more comparable to results of direct comparisons. So, we caution that evidence from indirect comparisons should not be combined with direct comparisons, if sufficient direct comparisons are available. It is also an implication for researches working on primary studies of diagnostic test accuracy: even diagnostic test can be evaluated without a competitor, but it is still valuable to perform a comparative study so that systematic reviewers can benefit from that.

Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.jclinepi.2014.10.005.

References

[1] Macaskill P, Gatsonis C, Deeks JJ, Harbord RM, Takwoingi Y. Chapter 10: analysing and presenting results. In: Deeks JJ, Bossuyt PM, Gatsonis C, editors. Cochrane handbook for systematic reviews of diagnostic test accuracy version 1.0. The Cochrane Collaboration, 2010. Available at: http://srdta.cochrane.org/.

[2] Song F, Xiong T, Parekh-Bhurke S, Loke YK, Sutton AJ, Eastwood AJ, et al. Inconsistency between direct and indirect comparisons of competing interventions: meta-epidemiological study. BMJ 2011;343:d4909.

[3] Takwoingi Y, Leeflang MMG, Deeks JJ. Empirical evidence of the importance of comparative studies of diagnostic test accuracy. Ann Intern Med 2013;158:544−54.

[4] Leeflang MMG, Di Nisio M, Rutjes AWS, Zwinderman AH, Bossuyt PMM. Adjusting for indirectness in comparative test accuracy meta-analyses. Cochrane Database of Systematic Reviews, supplement 2011;CD000003(Suppl):148.

[5] Broeze KA, Opmeer BC, van der Veen F, Bossuyt PM, Bhattacharya S, Mol BWJ. Individual patient data meta-analysis: a promising approach for evidence synthesis in reproductive medicine. Hum Reprod Update 2010;16(6):561−7.

[6] Broer SL, Dolleman M, Van Disseldorp J, Broeze KA, Opmeer BC, Aflatoonian A, et al. Prediction of an excessive response from patient characteristics and ovarian reserve tests and comparison in subgroups: an individual patient data meta-analysis. Fertil Steril 2013; 100:420−9.

[7] Jirge P. Ovarian reserve tests. J Hum Reprod Sci 2011;4(3):108−13.

[8] Broer SL, van Disseldorp J, Broeze KA, Dolleman M, Opmeer BC, Bossuyt P, et al, IMPORT study group. Added value of ovarian reserve testing on patient characteristics in the prediction of ovarian response and ongoing pregnancy: an individual patient data approach. Hum Reprod Update 2013;19(1):26−36.

[9] Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al, Standards for Reporting of Diagnostic Accuracy. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD Initiative. Clin Chem 2003;49:1−6.

[10] Broeze KA, Opmeer BC, van der Veen F, Bossuyt PM, Bhattacharya S, Mol BW. Individual patient data meta-analysis of diagnostic and prognostic studies in obstetrics, gynaecology and reproductive medicine. BMC Med Res Methodol 2009;9:22.

[11] Janes H, Pepe MS. Adjusting for covariates in studies of diagnostic, screening, or prognostic markers: an old concept in a new setting. Am J Epidemiol 2008;168:89−97.

[12] Karimollah HT. Methodological issues of confounding in analytical epidemiologic studies. Caspian J Intern Med 2012;3(3):488−95.

[13] Leisenring W, Pepe MS, Longton G. A marginal regression modelling framework for evaluating medical diagnostic tests. Stat Med 1997;16:1263−81.

[14] Zwinderman AH, Bossuyt PM. We should not pool diagnostic likelihood ratios in systematic reviews. Stat Med 2008;27:687−97.

[15] Pepe MS, Longton G, Janes H. Estimation and comparison of receiver operating characteristic curves. Stata J 2009;9(1):1−16.

[16] Janes H, Longton G, Pepe MS. Accommodating covariates in receiver operating characteristic analysis. Stata J 2009;9(1):17−39.

[17] Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 2011;12(1):1−8.

[18] Broer SL, Mol BW, Hendriks D, Broekmans FJM. The role of antimullerian hormone in prediction of outcome after IVF: comparison with the antral follicle count. Fertil Steril 2009;91:705−14.

[19] Schisterman EF, Perkins NJ, Liu A, Bondell H. Optimal cut-point and its corresponding Youden Index to discriminate individuals using pooled blood samples. Epidemiology 2005;16:73−81.

[20] Bancsi L, Broekmans FJM, Mol BWJ, Habbema JD, te Velde ER. Performance of basal follicle-stimulating hormone in the prediction of poor ovarian response and failure to become pregnant after in vitro fertilization: a meta-analysis. Fertil Steril 2003;79:1091−100.

[21] Hendriks DJ, Mol BW, Bancsi LSFJM, te Velde ER, Broekmans FJM. Antral follicle count in the prediction of poor ovarian response and pregnancy after in vitro fertilization: a meta-analysis and comparison with basal follicle-stimulating hormone level. Fertil Steril 2005;83:291−301.

[22] La Marca A, Sighinolfi G, Radi D, Argento C, Baraldi E, Artenisio AC, et al. Anti-Mullerian hormone (AMH) as a predictive marker in assisted reproductive technology (ART). Hum Reprod Update 2010;16(2):113−30.

[23] Song F, Harvey I, Lilford R. Adjusted indirect comparison may be less biased than direct comparison for evaluating new pharmaceutical interventions. J Clin Epidemiol 2008;61:455−63.

[24] Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. BMJ 2006; 332:1089−92.