Evaluating medical tests and biomarkers
Wang, J.

Citation for published version (APA):
Wang, J. (2017). Evaluating medical tests and biomarkers: In primary studies and systematic reviews

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

Download date: 30 Dec 2018
Using individual patient data to adjust for indirectness did not successfully remove the bias in this case of comparative test accuracy.

Junfeng Wang, Patrick Bossuyt, Ronald Geskus, Aeilko Zwinderman, Madeleine Dolleman, Simone Broer, Frank Broekmans, Ben Willem Mol, Mariska Leeflang, IMPORT Study group

Abstract

Objective

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

AMH and FSH differed significantly in sensitivity (-0.1563, p=0.04). AMH and AFC differed significantly in sensitivity (0.1465, p<0.01). AMH and AFC differed significantly in specificity (-0.0607, p=0.02). The 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.
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 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 study level data. This is the first time IPD is used to investigate and adjust for indirectness.

What is the implication, 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.

Keywords: Diagnostic test accuracy; Comparative meta-analysis; individual patient data; sensitivity and specificity; Receiver operating characteristic; Generalized estimating equation
4.1 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, while the second test is evaluated in different studies and different patients.

For various reasons, e.g. 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.

Inconsistency in the treatment effects between direct and indirect comparisons has previously been observed in systematic reviews of competing interventions [66]. This finding also applies to systematic reviews of diagnostic test accuracy. Takwoingi et al. 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 [2].

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 5 of them. In order 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. Leeflang’s results showed that adjusting for study features did not have much effect on removing the indirectness [67]. 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, i.e. 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 cut-off values [68].

The objective of this case study is 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.

4.2 Data

4.2.1 Data Acquisition

This IPD case study was facilitated by the EXPORT dataset 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) [69]. The dataset 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 [70]. AMH, AFC and FSH are three most widely used ORTs frequently used prior to IVF treatment to predict poor response to ovarian stimulation [71].

Patient characteristics, such as age, 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 [69]. 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 dataset, we can generate 3 pairwise comparisons between two tests: AMH vs FSH, AMH vs AFC, FSH vs AFC, which could make best use of the IPD dataset 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.

4.2.2 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, while the discriminatory power of continuous tests is usually measured with the area under a receiver operator characteristic (ROC) curve [72].

Many test results are continuous in nature, but classified as positive and negative,
thus in most of the diagnostic test accuracy studies, data are generally reported in a
dichotomous way, i.e. in 2 by 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.

4.3 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), 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, e.g. 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, since 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.
4.3.1 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 cut-off values to define diseased and non-diseased 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 dataset. So by adjusting reference standard, we can make sure that test accuracies in different primary studies are measured against the same reference standard and same cut-off 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 [73]. To make the pooling of data from primary studies more comparable, for each index test a single cut-off value should be defined and applied to all the patients in all studies. The general cut-off 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 cut-offs reported in the primary studies.

For continuous tests, there are no cut-offs but test results can differ between primary studies both in controls and cases. Janes and Pepe proposed a model to correct for the heterogeneity, by standardizing the test results for differences in the test result levels in controls between studies [74]. 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.
4.3.2 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 Type II adjustments.

When there are covariates associated with both test results and disease status, test accuracy may be over- or underestimated if these confounders are not considered in the design of diagnostic accuracy studies [75]. In direct comparisons, all the tests are evaluated based on 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, e.g. 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 [76]. 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 [75].

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 [77], so we compare dichotomous tests by their sensitivities and specificities. Pooling of sensitivity and specificity separately is not recommended since the paired nature of sensitivities and specificities from individual studies is ignored. A robust and commonly used approach is the construction of an SROC 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. (1997) allows comparing diagnostic tests with unbalanced data which contains both paired data (direct comparison) and
unpaired data (indirect comparison), and generalizes McNemar’s test for paired binary data [76].

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 showed that when confounding was present the overall ROC curve and AUC substantially differed from stratum specific ROC curve and AUC [74]. 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 (A\textsc{ROC}) curve [78, 79]. With this model, covariates that may influence the test accuracy could be statistically adjusted in the ROC analysis.

### 4.3.3 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 testtype, then the hypothesis \( H_0 : \beta_3^D = 0 \) (\( H_0 : \beta_3^D = 0 \)), where is the parameter of the interaction term in formula 4.1 (or formula 4.2) (see Appendix 1), 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 (non-parametric) 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 [80]. In Type II adjustments, A\textsc{ROC} was used as the measure of accuracy instead of AUC. Then we can see if there are inconsistencies between
comparative results from direct and indirect comparisons.

4.4 Results

4.4.1 Dataset

The final dataset only included women that 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 4.1 shows the number of patients in each pairwise comparison separated by the type of comparisons.

<table>
<thead>
<tr>
<th></th>
<th>FSH vs AFC</th>
<th>AMH vs AFC</th>
<th>FSH vs AMH</th>
</tr>
</thead>
<tbody>
<tr>
<td>FSH</td>
<td>2248</td>
<td>1024</td>
<td>1747</td>
</tr>
<tr>
<td>AFC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AMH</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AFC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FSH</td>
<td>2108</td>
<td>252</td>
<td>867</td>
</tr>
<tr>
<td>AMH</td>
<td>1476</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AFC</td>
<td>2609</td>
<td>144</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.1: Number of patients in each pairwise comparison

4.4.2 Reference Standard

Different cut-off points were used in primary studies to define ‘poor response’. Since in our IPD dataset 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 4 or less oocytes at follicle [81].

4.4.3 Dichotomous Tests

In this study, we defined the single cut-off value for each ORT in all studies by maximizing the Youden’s index [82] (Sensitivity + Specificity - 1) based on the dataset as well as consulting with recently published systematic reviews of each ORT [83–85]. 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 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 cut-offs reported in the systematic reviews of these ORTs [83–85].

Comparisons between the tests sensitivities are based on the marginal regression model in formula 4.1 (See Appendix 1), 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 4.2, Appendix 1) 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 4.2.

Table 4.2 shows the parameter estimates and 95% confidence intervals (in parentheses). For sensitivity, the differences between AMH versus FSH (-0.1563, p=0.04) and AMH versus AFC (0.1465, p<0.01) in direct and indirect comparisons are significant. For specificities, the difference between AMH versus 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 $\beta^3$ 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.
Table 4.2: Number of patients in each pairwise comparison

<table>
<thead>
<tr>
<th>Analysis 1</th>
<th>Analysis 2</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMH vs FSH</td>
<td>0.0469 (0.0208, 0.0748)</td>
<td>0.0194 (0.0158, 0.0230)</td>
<td>0.0043 (0.0034, 0.0053)</td>
</tr>
<tr>
<td>AMH vs AFC</td>
<td>0.0439 (0.0209, 0.0676)</td>
<td>0.0194 (0.0158, 0.0230)</td>
<td>0.0043 (0.0034, 0.0053)</td>
</tr>
<tr>
<td>FSH vs AFC</td>
<td>0.0469 (0.0208, 0.0748)</td>
<td>0.0194 (0.0158, 0.0230)</td>
<td>0.0043 (0.0034, 0.0053)</td>
</tr>
</tbody>
</table>

Note: Numbers in parentheses are 95% confidence intervals of parameter estimates. Significant differences between direct and indirect comparisons are in bold.
4.4.4 Continuous Tests

The ROC curves and AUCs after adjusting for reference standard and standardizing the test results (Type I adjustments) are shown in Figure 4.1 and Table 4.3. 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 were 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 to 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 due to 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>
<thead>
<tr>
<th>Table 4.3: Comparisons of AUCs in each pairwise comparison</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. AMH vs FSH</td>
</tr>
<tr>
<td>Direct comparison</td>
</tr>
<tr>
<td>Indirect comparison</td>
</tr>
<tr>
<td>2. AFC vs FSH</td>
</tr>
<tr>
<td>Direct comparison</td>
</tr>
<tr>
<td>Indirect comparison</td>
</tr>
<tr>
<td>3. AFC vs AMH</td>
</tr>
<tr>
<td>Direct comparison</td>
</tr>
<tr>
<td>Indirect comparison</td>
</tr>
</tbody>
</table>

Figure 4.2 and Table 4.4 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.
4.5 Discussion

In this IPD case study, we proposed two types of adjustments to correct for the effect of indirectness in comparative systematic reviews of diagnostic test accuracy studies. Type I adjustments were focused on threshold effect and reference standard issues, while 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.

**Figure 4.1:** ROC curves of pairwise comparisons in direct and indirect comparisons after adjustment for heterogeneity
Table 4.4: Comparisons of AUCs in each pairwise comparison after covariate-adjustment of age

<table>
<thead>
<tr>
<th>Pair</th>
<th>AMH</th>
<th>FSH</th>
<th>Diff</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. AMH vs FSH</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Direct comparison</td>
<td>0.7669 (0.74, 0.79)</td>
<td>0.6661 (0.63, 0.70)</td>
<td>0.1008</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Indirect comparison</td>
<td>0.7374 (0.64, 0.83)</td>
<td>0.6095 (0.58, 0.64)</td>
<td>0.1279</td>
<td>0.0135</td>
</tr>
<tr>
<td>2. AFC vs FSH</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Direct comparison</td>
<td>0.7193 (0.69, 0.74)</td>
<td>0.6424 (0.61, 0.67)</td>
<td>0.0769</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Indirect comparison</td>
<td>0.7051 (0.63, 0.78)</td>
<td>0.6375 (0.60, 0.67)</td>
<td>0.0676</td>
<td>0.0985</td>
</tr>
<tr>
<td>3. AFC vs AMH</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Direct comparison</td>
<td>0.7207 (0.68, 0.76)</td>
<td>0.7332 (0.70, 0.77)</td>
<td>-0.0125</td>
<td>0.4797</td>
</tr>
<tr>
<td>Indirect comparison</td>
<td>0.7127 (0.68, 0.75)</td>
<td>0.7949 (0.76, 0.83)</td>
<td>-0.0822</td>
<td>0.0006</td>
</tr>
</tbody>
</table>

Analyses were performed with both dichotomous tests and continuous tests. Since most of the diagnostic test accuracy 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, since 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.

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. found that indirect comparison with adjustment may be less biased
Figure 4.2: ROC curves of pairwise comparisons in direct and indirect comparisons after covariate-adjustment of age

than direct comparison [86]. 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. There are also other limitations of our research. First, for certain types of diagnostic tests, such as CT and MRI 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 dataset, 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 dataset or even not known. Fourth, methods for doing DTA IPD meta-analysis vary among systematic reviews, due to the different aims of the reviews and data available. Since 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 vice versa between two tests. This information can be observed from the ROC curves. Last but not least, this study is only a case study. Since IPD is seldom available, especially for which comparing two or more tests and contains both direct and indirect comparisons, we have only one dataset 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 [87]. 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 difference decision in clinical practice. For example, the difference found in logit sensitivity between AMH and FSH (Table 4.2) 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, while 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 diagnostic test accuracy studies is needed. Systematic reviews will also benefit from better design of primary studies. Comparative studies should be encouraged in DTA studies.

4.6 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.
Acknowledgments

This project was supported by The Netherlands Organization for Scientific Research (NWO); project 916.10.034.
Appendix

Models for dichotomous tests

The model for sensitivity has the form:

\[
\logit P[Y_{ij} = 1 | D_i = 1, X_{ij}, Z_i] = \beta^0_\bar{D} + \beta^1_\bar{D}X_{ij} + \beta^2_\bar{D}Z_i + \beta^3_\bar{D}X_{ij}Z_i \quad (4.1)
\]

and for 1-specificity it has the form:

\[
\logit P[Y_{ij} = 1 | D_i = 0, X_{ij}, Z_i] = \beta^0_D + \beta^1_DX_{ij} + \beta^2_DZ_i + \beta^3_DX_{ij}Z_i \quad (4.2)
\]

where \( i \) stands for individual patient \( (i = 1, 2, 3, \ldots) \), \( j \) stands for index test, \( Y \), \( D \) and \( X \) denote binary variables that indicate index test result \( (Y=1 \text{ for positive}) \), disease status \( (D=1 \text{ for present}) \) and test type \( (X=1 \text{ for test A and } X=0 \text{ for test B in each pairwise comparison between test A and B}) \); \( Z \) is the covariate that may influence sensitivity and specificity.

Models for continuous tests

Calculating standardized test result:

\[
pv_{ij}(x) = \Pr(X_{ij} < x | D = 0) \quad (4.3)
\]

where \( X_{ij} \) is from the distribution of index test \( i \) of the control population in study \( j \).

Covariate adjustment:

\( AROC \) is a stratified measure of diagnostic test performance which is defined as:

\[
AROC(f) = \Pr(1 - pv_{DZ} \leq f) \quad (4.4)
\]
where \( pv \) stands for percentile value as in formula 3, \( pv_{DZ} \) represents the patient who has the disease with the covariate value \( Z \) standardized with the control population with the same covariate value \( Z \). If we assume covariate \( Z \) to act linearly on test results \( Y \), can be calculated as:

\[
 pv_{DZ} = \hat{F}\{(Y - \hat{\beta}_0 - \hat{\beta}_1 Z)/\hat{\sigma}\} \tag{4.5}
\]

where \( \hat{\beta}_0, \hat{\beta}_1 \) and \( \hat{\sigma} \) are estimates from the linear model, and \( \hat{F} \) is the c.d.f of the error distribution estimated empirically.