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