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.
Summary and conclusions
This thesis is focused on the evaluation of medical tests and biomarkers, in both primary studies and in systematic reviews and meta-analyses. It presents a number of test evaluations, and methodological developments to improve the methods for such evaluations.

Many clinical questions around medical tests are comparative, or can be addressed as a comparative question [87]. In Chapter 2, we present an example of a systematic review about diagnostic test accuracy addressing such a comparative question. We reviewed the accuracy of three index tests in detecting clinically suspected scaphoid fractures: computed tomography, magnetic resonance imaging and bone scintigraphy. The systematic review included studies that evaluated only one of the mentioned tests, and studies that directly compared two tests in the same patients and against the same reference standard. From previous research, we know that these comparative studies may provide a different answer to the question which test is most accurate compared to single test accuracy studies [2]. Following the current Cochrane recommendations [1], meta-analyses for single tests were performed with the bivariate model and HSROC model and by adding index test as a covariate in the HSROC model for pairwise comparison. In the pairwise comparison, the main analyses were based on all studies (both single test studies and comparative studies), and a sensitivity analysis including only comparative studies was done to assess whether the comparative-only study meta-analysis give a different answer than the main analysis. The meta-analyses showed that bone scintigraphy had a slightly lower specificity but a much higher sensitivity than CT and MRI leading to a significantly higher overall accuracy of bone scintigraphy when compared with CT and MRI, while no evidence was found for a difference in accuracy between CT and MRI.

Although we assume that, in general, the comparative-only meta-analysis provides a less biased result, there is currently limited evidence for that assumption. The statistical analyses for the review reported in Chapter 2 invited us to think about the following questions: are the current statistical method for comparative meta-analysis appropriate? When we have both direct comparisons (head to head comparison
between two or more index tests) and indirect comparisons (index tests evaluated in
different primary studies), what is the proper way to synthesize the evidence? Which
performance measure should we report as the results, summary point, summary ROC
curve or the area under the summary ROC curve? These questions were addressed
in the subsequent chapters.

One of the solutions for meta-analysing both comparative studies and single test
studies, may be to develop a methodology that takes into account the advantages of
the comparative studies to make inferences about the comparative accuracy that can
be used for the analyses containing single test studies. In Chapter 3, we developed
two approaches for comparative meta-analysis of diagnostic tests, taking into account
the characteristics of three different study types: single test studies; comparative
studies providing the same data structure as two (or more) single test studies; and
comparative studies providing data about the correlation between the two index tests.
The absolute accuracy approach (arm based) first estimates the mean diagnostic
accuracy statistics for each test and then calculates the difference between diagnostic
tests, while the relative difference approach (contrast based) first calculates the
difference in accuracy between diagnostic tests within each study and then estimates
the mean difference between the tests. In both approaches, we modeled the test
outcomes directly with a multinomial distribution, and simulations showed that both
approaches yielded comparable results when only head-to-head comparisons are
used for meta-analysis. The model can be extended to include covariates, or to
include more than two tests.

Although the advanced statistical methods may be a solution to the question
how comparative meta-analysis should be conducted, the co-existence of primary
studies comparing two or more index tests (evidence from direct comparison) and
those evaluating only one index test (evidence from indirect comparison) may still
jeopardize the reliability of the results. The problem is twofold: (1) both the single
test evaluations and the comparative evaluations may be biased, and (2) the single
test studies evaluating test A may have been done in a different patient population
than the studies evaluating test B.

We then focused on the second problem. We know that single test studies can provide different answers than comparative studies, but it is not always clear what drives these differences. Previous research showed that, in some cases, adjusting for patient characteristics may attenuate the differences between single test and comparative studies [191].

In Chapter 4, we investigated the difference between results from meta-analysis of direct comparisons and from indirect comparisons, using individual patient data (IPD). First, we confirmed our previous findings about the existence of a bias in indirect comparisons and we considered this difference to be caused by bias from these indirect comparisons. There are two main sources of bias in indirect comparisons: heterogeneity between studies, which may be generated by, for example, different reference standards or different thresholds in primary studies; and differences in baseline characteristics, which can lead to differences in test performance. Accordingly, we proposed two types of adjustments to correct for these biases: Type I adjustments, focused on threshold effect and reference standard issues, and Type II adjustments, which additionally focus on patient characteristics (e.g., age). 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.

The issue of indirectness in the comparison of diagnostic tests is more complex than we expected and addressing it was found to require a different approach. First of all, we do not know sufficiently how comparative accuracy may be biased, or influenced by other factors. We also know that relying on single test studies may lead to bias in comparative meta-analyses, but we have no empirical data yet on the potential bias in comparative studies and paired data. For example, certain patient groups may have been excluded from comparative studies, because receiving two tests was not necessary for them or even harmful. However, if these patients are the patients who will be tested using either test A or test B, then the comparison
will suffer from selection bias. Simply adding covariates into the meta-regression will not necessarily solve this problem, because we need to realize that covariates may influence the test performance as confounders or effect modifiers. Clever thinking about how a comparison between two tests may be influenced and what factors may play a role there, is therefore needed.

Chapter 5 dealt with another aspect of diagnostic accuracy meta-analysis: the performance measure. Although Cochrane advises against using the Area Under the ROC curve (AUROC), in some areas of medicine, the AUROC is the main accuracy measure reported in primary studies. Systematic reviews meta-analyzing this type of studies therefore feel the urge to report the Area under the summary ROC curve (AUSROC), based on, for example the recommended bivariate model or HSROC model. We investigated the use of the AUSROC as an overall performance measure in meta-analysis of diagnostic accuracy studies and we investigated the validity of this measure. In our simulation study, we showed that it is virtually impossible to estimate the correct AUSROC from the current meta-analysis of 2-by-2 tables. The direction and the amount of bias may depend on the way the threshold was selected in the primary studies. In practice, the AUSROC should therefore not be used as a test accuracy measure in systematic reviews of diagnostic accuracy.

The AUSROC is mainly used for tests with a continuous test result. For these tests, Cochrane advises to use the HSROC model. This model provides the reader with an overall measure for accuracy, with a measure for the threshold at which the test is used on average and the shape of the curve. Although AUROC is difficult to interpret, clinicians in certain areas of medicine seem to be more comfortable in interpreting AUROC than other measures of test accuracy. Further research may therefore also this preference, and where, how and when alternative summary measures of test accuracy may be acceptable for them.

We did not investigate the value of the AUSROC as a measure of comparative meta-analyses. One of the problems, well known form the AUC in single test evaluations, is that similar AUSROCs can be based on very different summary ROC
curves, complicating comparisons between AUSROCs. If one test has a lower AUSROC, but the range of the summary ROC curve where the test is normally used indicates that the test has a higher sensitivity or specificity than the test operating range of the test with the larger AUSROC, then the test with a lower AUSROC may still be the clinically more usually.

There is still much room for improvement in the methodology for (comparative) diagnostic test accuracy meta-analyses.

The basis of the models developed in Chapter 3 was a multi-dimensional distribution to model the (joint) true positives and true negatives of two or more index tests and their correlations. Usually this multi-dimensional distribution will have a very complex variance-covariance matrix. Estimating the parameters and the matrix can be challenging. Some researchers have looked at the question from the other side [192–194]. They first defined a unified framework for network meta-analysis of any N index tests. These frameworks provided flexibility to include all the relevant evidence of test accuracy into one meta-analysis, but they more or less offer only conceptual solutions. We believe there are two routes to solve the problem of comparative meta-analysis of DTA studies: to start either from the statistical methods that are currently used in DTA meta-analysis or from the framework of network meta-analysis for interventions. Although each approach takes a different path towards the very same goal, the origins of the respective methods make them different in essence. Further research is needed to compare these different strategies and to assess the appropriateness of each.

In Chapter 4, we used IPD meta-analysis to investigate and adjust the bias from indirect comparisons. In the statistical analysis, we chose a GEE model to account for the correlation between index tests within patients. However, there are alternative approaches for IPD meta-analysis of DTA studies, and it is not clear yet which approach is preferred. It is generally assumed that adjusting for the fact that IPD coming from different studies by using a random intercept logistic regression model is better than just pooling all IPD as if they came from one big study. However, we
learned from the research reported in Chapter 5 that the latter approach introduces less bias than the first approach. The ongoing project about review of statistical methods for IPD meta-analysis of DTA studies will give an answer to this question [195].

It is clear that, at this point, we need a new research scheme for comparative meta-analyses. First of all, we need more comparative primary studies that also report paired data for the index tests. Second, we need more information about the mechanisms leading to bias in primary comparative studies and in comparative meta-analyses. Third, we need effective ways to remove any bias from single test evaluation studies.

The second part of the thesis dealt with evaluation of the prognostic performance of biomarkers and clinical prediction models that include such markers. The key word for Part 2 of the thesis is the C-statistic, especially in survival analysis. When we have survival data and baseline biomarkers values, we can use both the overall C-statistic (Harrell’s) and the time-dependent C-statistic (e.g. Uno’s) to evaluate the prognostic value of the marker. We can also use the C-statistic as the criterion for variable selection whenever we develop a multi-variable prognostic model.

In Chapter 6, we evaluated the prognostic value of alkaline phosphatase (ALP) both at diagnosis of primary sclerosing cholangitis (PSC) and one year later. Instead of looking at the hazard ratio in a Cox model, we used the time-dependent C-statistic to assess both the short-term and long-term predictive value of ALP. Performances of ALP at T0, T1 and the relative change were compared over follow-up time, and the optimal threshold was determined based on the overall C-statistic. We concluded that ALP at T1 has a better prognostic performance than ALP at T0 by visually comparing the time dependent C-statistics, but no formal statistical test was performed, since both evaluations were based on the same patients. So the time dependent C-statistic only provided an indication of prognostic value over time. If no specific time point is of interest, the comparison should not be based on this time dependent C-statistic.

In the project reported in Chapter 7, we continue working on the prognosis of
PSC. We aimed to develop a prognostic model with multiple biomarkers and other patient characteristics. These biomarkers had large amount of missing values at time of diagnosis, and some of them were measured repeatedly during follow-up. To deal with the missing values at baseline, multiple imputation technique was employed. We chose the following imputation models for different types of variables. For continuous time-fixed variables predictive mean matching was used; for binary variables logistic regression was used, while for variables that could change over time (AST, ALT, ALP and total bilirubin) a two-level linear model was used. Even for the patients for which we had the baseline biomarker values in our dataset (values that were obtained within three months (+/- 3 months) from diagnosis) in our dataset, we still used the predicted value at T0 as baseline value in our model rather than the real observed value. We believe the predicted (imputed) biomarker value, which contains information from follow-up measurements is more suitable for studying the effects of those markers. The multiple imputed dataset brings new challenges in variable selection: the selected variables may differ per imputed data set, which can be problematic when we want to average over the parameter estimates from the 20 imputed data sets. The solution we chose was stacking all datasets to create a long combined dataset and using Lasso to shrink the parameters of some variables to 0, after which these were excluded from the model. Lasso’s penalty parameter 'lambda' was determined based on the discriminative power of the model using Harrell’s C-statistic, with the criterion that the resulting model has as few predictors as possible while still yielding a C-statistic that is not more than 10% below the optimal one. The internal validation was integrated into the model derivation step to adjust for the optimism in the C-statistic, and then the model was externally validated with an independent cohort from Oxford, UK. One additional step of re-calibration was taken to compensate for over-shrinkage. The re-calibration was performed on the prognostic Index derived from Lasso but not on the selected variables, since we would prefer to keep the relative importance of all variables and avoid over-fitting and inflation of parameters.
An online risk calculator is available for both clinicians and patients to predict the outcome (probability of survival at any year after diagnosis). The calculator plays a role similar to that of a nomogram, but it is probably more user-friendly. One of the reasons why we need the variable selection step in model derivation is that a model with too many variables may not be easy for clinical use. Implementation of the calculator can easily solve this problem, as it allows the prognostic model to have more variables. Future model development may suffer less from limitations in the number of variables, at least from the clinical use perspective, unless there are other concerns, for example, when markers are very expensive to measure.

In the project reported in Chapter 8, the similar analyses similar to those of Chapter 6 were performed for a different disease, and for several biomarkers: the tubular damage markers urinary neutrophil gelatinase-associated lipocalin (NGAL), kidney injury molecule-1 (KIM-1), glomerular damage markers urinary albumin and total protein. The prognostic value of these markers on CKD stage transition was evaluated with time dependent C-statistics. The new challenge in this study was adjusting the baseline eGFR. The Cox model with baseline eGFR and eGFR-adjusted overall C-statistic was used, but a covariate adjusted time-dependent C-statistic was not available which may be a topic for further research. The reason why we need this covariate adjusted C-statistic is that biomarker values of patients in different stage or with different baseline eGFR can not be compared directly. A Stage 1 patient may have a higher probability of transition than a patient in Stage 2 with the same marker value. A patient with baseline eGFR near the boundary is more likely to have a transition.

Although Part 2 does not explicitly contain comparison between biomarkers, the variable selection step in the derivation of a multiple variable prognostic model provides some indication of such a comparison. Traditionally, the importance of the predictors (in this case, biomarkers) can be expressed by the standardized coefficients in the regression model. In our study, we found Lasso not only performed the variable selection but also provided a ranking of the strength of the predictors by the order of
shrinking them to 0.

The evaluation of medical tests still lags behind the evaluation of pharmaceuticals and of other medical interventions. The methods for assessing the performance and effectiveness of biomarkers, tests and models are less well developed, both for primary studies and for evidence synthesis based on systematic reviews. As becomes clear from this summary, the research reported in this thesis did not solve all problems, but it offers methodological advances and applications of state-of-the-art statistical techniques that can inspire and help other researchers. We are confident that future developments will further strengthen the concepts and techniques for the evaluation of medical tests, thereby providing the evidence base needed for rational decision-making about medical tests.