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.
Chapter 1

Introduction

Junfeng Wang, Mariska Leeflang
Medical tests and biomarkers are used to discriminate between people who do have a certain target condition or disease (diagnosis) and those who do not. Tests may also be used to discriminate between people who will have a certain event (prediction or prognosis) and those who will not. Both tests can be used on their own or in combination. Many tests can also be used in a dichotomous way, other produce results along a continuum.

This thesis addresses methodological problems in the evaluation of tests and biomarkers, in quite distinct types of research. The first part addresses systematic reviews and meta-analysis of diagnostic accuracy: discrimination between diseased and non-diseased (cross-sectional design; dichotomous test use). The second part presents primary studies of prognostic factors and clinical prediction rules: the probability of developing an event (continuous test result) in studies that include a time element. Both parts focus on discrimination between patient groups and the best ways to assess and express the discriminating ability of a test or model.

1.1 Part 1: Comparative meta-analysis of diagnostic accuracy

Although most diagnostic accuracy studies investigate the diagnostic accuracy (e.g., sensitivity and specificity) of one test of interest, many clinically relevant questions are comparative. Clinicians not only want to know the accuracy of ultrasound for detecting appendicitis, for example, but also whether using the ultrasound adds information to using only clinical signs and symptoms. Or whether ultrasound is at least as accurate as MRI scanning. Another example may be biomarkers used to detect cancer: which of the available biomarker is the most accurate one?

Clinicians looking for answer to these questions, usually turn to systematic reviews and meta-analyses of diagnostic test accuracy. Recently, more attention has been given to comparative meta-analysis of diagnostic accuracy. Such a meta-analysis
answers the question whether one test has a higher accuracy - i.e. higher sensitivity or higher specificity, or bothy - than the other. These meta-analyses may have used the recommended meta-regression models for diagnostic meta-analysis and added test-type as a covariate to those models [1]. They provide the reader with accuracy measures (diagnostic odds ratios, sensitivity, specificity) for each of the test-types included and may have tested for a statistically significant difference between the test-types. However, this approach has some limitations and disadvantages.

First of all, the meta-analysis may have included studies that evaluated only one of the tests of interest against a reference standard. It is possible that the tests of interest were evaluated in different studies, at different times, in different populations and even against different reference standards. That means that if test A was mainly analysed in studies done in primary care and test B was mainly analysed in secondary care, then any differences in accuracy can be caused by setting rather than by test-type. Meta-analyses including these single test designs, may lead to bias compared to meta-analyses based on studies comparing both tests against the same reference standard and in the same patients [2].

It therefore makes sense to include studies comparing both (or all) tests of interest in the review, performed in the same population and against the same reference standard. A second disadvantage of the methodology described earlier is that it does not take advantage of any direct comparisons, if they have been performed. Such comparative studies could provide information about the disagreements between the tests under evaluation; this knowledge may increase the power of the analyses to detect a differences between the tests.

Methods have been developed to compare the accuracy of multiple diagnostic tests in a meta-analysis. However, none of these methods has actually been tried out on more than two tests and most methods are not capable of properly dealing with primary studies that do not provide data to assess the agreement between the index tests. We have therefore improved this methodology and propose different sampling strategies for different study types.
In Chapter 2, we first start with one example of a systematic review and comparative meta-analysis of diagnostic accuracy of three tests. We added test type as covariate into the HSROC model and tested the statistical significance of the covariate effects on the test accuracy. As this method has disadvantages and statistical method for comparative meta-analysis are behind in methodological development, we aimed to develop a new statistical method for comparative meta-analysis of multiple diagnostic tests in Chapter 3. We proposed two approaches either assessed the difference between accuracies after the accuracy per test was meta-analysed (the absolute, or arm-based approach) or it meta-analysed the difference in accuracy for each study (the relative, or contrast-based approach). Even though the development of proper statistical methods can facilitate the comparative meta-analysis, the primary studies may not be conducted in a comparative design. Single test studies may provide bias to the meta-analysis, so we investigated the possibility to adjust for bias from single test accuracy studies, with individual patient data (Chapter 4). In Chapter 5 we assessed a method for test comparisons that is discouraged, but still widely used. The area under the ROC curve (AUROC) may be comparable between tests, while their actual performance is quite different. On top of this (and other) conceptual reason for not recommending the AUROC, we also provide statistical evidence why the area under the summary ROC curve should not be used as an overall accuracy measure in a meta-analysis of diagnostic tests.

1.2 Part 2: Prognostic performance of biomarkers and clinical predictive models

Just as we sometimes want to compare and rank diagnostic tests by their accuracy, we may also be interested in using the most accurate biomarkers, or combinations thereof to make a distinction between patients. Such a distinction can either be based on a ranking of the patients’ probability of having a certain disease (diagnosis), the patients’ severity of the disease (staging), the patients’ risk of getting an event
(prognosis), or the likelihood of a treatment response or treatment benefit (treatment selection). In all these cases, biomarkers and often combined with other patient characteristics into multivariable predictive models.

Prognostic models with biomarkers can also be seen as medical tests. The differences between these and the diagnostic accuracy paradigm are (1) that the models usually provide a continuous test result, which is then translated to a (continuous) probability of a certain event happening; and (2) that there can be a time-element involved in a prognostic model. Like diagnostic tests, prognostic models can also be assessed based on their discriminatory ability.

In Chapter 6, we evaluated a single biomarker (ALP) for its prognostic performance over time for primary sclerosing cholangitis (PSC). The biomarker was evaluated in a survival analysis setting. The prognostic performance of ALP measured at T0 was compared with T1 based on time dependent C-statistic.

In Chapter 7, we continued working on the prognosis of PSC, and we derived a prognostic model with a number of potentially relevant biomarkers and patient characteristics. This also concerns a comparison between markers: only the markers with acceptable prognostic value will be included in the model while other markers with weak prognostic value will be excluded.

In Chapter 8, we used a methodology similar to the one in Chapter 6 to evaluate the prognostic performance of several biomarkers for CKD stage transition. This time, we did not compare the biomarker performance at different time points, but evaluated the performance in patients in different CKD stages. Since the baseline eGFR plays an important role in the prognosis, we also calculated the eGFR adjusted C-statistic for all the markers.

### 1.3 C-statistic and its variants

The C-statistic or concordance C, has been widely used as a measure of discriminatory power of continuous tests, biomarkers or clinical predictive models. This
C-statistic, which is equivalent to the area under the ROC curve (AUC), was originally used for evaluating logistic regression models or other classification models to discriminate between patients based on their probability of disease at a certain time point. It has since been extended for several other purposes, such as the covariate adjusted C-statistic (used in Chapter 4 and Chapter 8), the C-statistic from summary ROC curves (Chapter 5), overall C-statistic in time to event data (used in Chapter 6 and Chapter 7) or the time-dependent C-statistic (used in Chapter 6 and Chapter 8).

The time-dependent C-statistic is relevant for survival analysis, where the time component is introduced into the calculation. In such a situation, the C-statistic reflects the prognostic value of the biomarker for any time point in the follow-up. Like the conventional C-statistic, the time dependent C-statistic can be calculated from paired (time dependent) sensitivity and specificity. There are three methods to estimate time dependent sensitivity and specificity: Cumulative/Dynamic, Incident/Dynamic and Incident/Static.

Covariates can influence the biomarker’s performance. When we evaluate or compare biomarkers, we should take the effect of covariates into account. Covariate adjustment is applicable for both cross sectional setting and survival setting. There are established approaches for overall C-statistic in survival analysis, and there are also some attempts to adjust for the covariate effect for all the time points (time dependent C-statistic).

1.4 Variable selection, shrinkage, and re-calibration

In model derivation we usually have a lot of candidate variables that may be potentially relevant in predicting the outcome. In the final prognostic model, however, we usually do not want to have such a complex model, which is often not easy to use, to understand, or to interpret to other stakeholders.

There are many techniques for variable selection; selection can be based on model fitting, the significance of coefficients, and the predictive power. Stepwise
regression has been widely used in variable selection, but it has many drawbacks [3]. Lasso has gained popularity since it can work for correlated data; the shrinkage can prevent over-fitting and may help with the generalization of the model.

After assessing the discriminatory power of a marker or a prognostic model, calibration accuracy is also an important feature for model use. If a model has good discrimination but poor calibration, a re-calibration of the model is needed. Usually, a shrinkage model can yield good discrimination, but prediction may be distorted and not corresponding to the observed outcome, especially when we need a large penalty to make more variables to be excluded. Thus, in our modeling practice in Chapter 7, we added an additional re-calibration step which we believe is necessary.

1.5 Internal and external validation

After model derivation, the actual performance of model should be evaluated for its validation. This can be done internally (with the derivation data, using bootstrap or cross validation techniques) or externally (with independent data, which can be either out of sample, or out of time). It is believed that external validation is more meaningful than internal validation. So we focused on external validation performance, and integrated internal validation into the model derivation step. In other words, we selected the model with the best performance in internal validation rather than the model that performs best in derivation, and then use external dataset to test model performance.