Beyond diagnostic accuracy. Applying and extending methods for diagnostic test research
Glas, A.S.
The current demand for evidence-based practice generates the need for proper evaluations of the effectiveness of diagnostic and therapeutic interventions. Diagnostic effectiveness reflects the performance of diagnostic tests under ordinarily, rather than ideal conditions. Within diagnostic effectiveness studies different levels of test evaluations can be distinguished, ranging from the technical development up to cost-effectiveness.

The most frequently performed type of study is the evaluation of a test's diagnostic accuracy. In such evaluations the results of a new test are compared with the results of a reference standard - formerly called the gold standard - which is the best available method to reveal the true disease state of the patients. When both the true disease state and the outcome of the test are dichotomous, to be classified as positive or negative, the results can be structured in a 2x2 contingency table.

Several measures of diagnostic accuracy can be calculated from such a table. The best known are the sensitivity and specificity, which stand for the proportion of positive test results within patients with the target disease and the proportion of negative test results within patients without the target disease, respectively. Other measures of diagnostic accuracy are the positive and negative predictive values, the likelihood ratios of a positive and negative test result, the overall accuracy and the Youden index. When the test outcome is continuous, while the true disease state is dichotomous, the likelihood ratios of the respective test results, the receiver operating characteristic curve and the area under this curve can be used. All these measures express the correspondence between the results of the test and the outcome of the reference standard, and can be seen as expressions of how well the test is able to distinguish between patients with the target disease from those without it.

This thesis explores the characteristics of a measure beyond the above mentioned ones: the diagnostic odds ratio (chapter one). It also provides more detailed information on the definition, interpretation and use of the outcome measures relative to the odds ratio. Yet, the reduction of the full test information into a 2x2 table, and the subsequent calculation of these measures may not be straightforward. Chapter two illustrates a number of pitfalls in studies that evaluate tests that can detect and localise specific lesions and try to force the resulting data in a 2x2 table. We present an suitable method to arrive at valid sensitivity and specificity. This method is illustrated in chapter four. In the study reported there, sensitivity and specificity were estimated for CT colonography in the detection of polyps in the colon at different levels of radiation dose.

When one wants to arrive at valid and precise estimates of a test's worth, one
can review the literature for available evidence on the accuracy of a diagnostic test. If one uses an explicit and complete search and includes criteria for evaluating the potential for bias in the studies found, the exercise is called a systematic review. A researcher who wants to do a systematic review aims at identifying all studies that have compared the test of interest with the reference standard. If enough comparable studies are available, a meta-analysis can be performed to arrive at summary measures of test performance.

A well known difficulty in meta-analysis of diagnostic tests is the existence of considerable variation between the results from different studies. A part of that variability can be attributed to differences in the threshold used to classify a test result as positive or negative. Simple averaging the results, with only adjusting for the sample size, may then yield invalid results. The most widely used class of methods for meta-analysis of diagnostic accuracy studies is based on the concept of the summary receiver operating characteristic (SROC) curve, which can take this threshold effect into account. This curve is a global measure of the accuracy of the test, which however has an ambiguous interpretation, and which does not allow for straightforward calculation of standard errors. Chapter three introduces a bivariate approach as an alternative to estimate summary values of the well established measures sensitivity and specificity. This method can be extended to a meta-regression technique to explore sources of between-study variability separately on sensitivity and specificity. Chapter five contains an application of this bivariate approach to the meta-analysis of diagnostic tests in a systematic review of urine based tumor markers in the primary diagnosis of bladder cancer.

All measures of diagnostic accuracy share a disadvantage, they do not indicate what the test’s added value is relative to what is already known about the patient. The added value of a diagnostic test depends by definition on the information available before testing, in particular on the likelihood of (non)disease. Chapter six of this thesis reports on a study of the added value of the dobutamine stress echocardiography in the diagnosis of cardiac disease in patients at known low risk. This test has been shown to be associated with the cardiac prognosis of patients discharged from an emergency department. At discharge, physicians will have a prognostic idea about these patients. We invited these physicians to express their prognostic guess as (subjective) probability estimates without knowing the test result and evaluated the contribution of dobutamine stress echocardiography. The test results can be linked to the pre-test probability estimates using likelihood ratios and Bayes theorem.

It is unclear as to what extent subjective probability expressions assigned by physicians can be relied on. Chapter seven presents a comparison of subjective
probabilities and two formal diagnostic decision rules in the diagnosis of ankle fracture in patients with a fresh ankle sprain. In chapter eight, an evaluation is described of physicians’ subjective judgment in interpreting the ventilation/perfusion lung scan in the diagnosis of pulmonary embolism. We used the subjective pre-test and post-test probability estimates of the lung scan to derive subjective likelihood ratios for the respective test results. These subjective likelihood ratios were then compared to the objective likelihood ratios of the lung scan, as obtained from the comparison of the test results with the outcome of the reference standard.

In general the difference between pre-test and post-test probabilities can be used to characterize the information content of a diagnostic test. Chapter nine presents an example of how the information content can be used to compare computer tomography with magnetic resonance imaging in the diagnosis of patients with lumbar radicular syndrome.

Tests have to be properly evaluated to see whether or not they are able to improve or maintain the health status of patients. This thesis is based on both existing methods and an extension of the methodology used to evaluate diagnostic tests. In the final chapter we summarize the state of the art of diagnostic methodology and provide indications for future research.

References