Beyond diagnostic accuracy. Applying and extending methods for diagnostic test research
Glas, A.S.

Citation for published version (APA):

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.
General Discussion

Methodological relevance of this thesis

This thesis applies and extends methodological tools that range from the field of diagnostic accuracy to the field of diagnostic impact studies. Diagnostic accuracy studies deal with how good a test distinguishes the sick from the well. Diagnostic impact refers to the effects of a test-result on the physicians' confidence in a particular diagnosis.\(^1\)

With regard to the diagnostic accuracy we encouraged the use and interpretation of the diagnostic odds ratio in chapter 1. By adding this test measure of discriminating performance to the set of available measures we feel a comprehensive toolkit of accuracy measures is available to address the many study questions possible when evaluating the discriminating performance of diagnostic tests.

The most frequently used pair of outcome measures is the combination of sensitivity and specificity. Yet estimating these measures is not clear cut in all situations. In chapter 2 a method is proposed to assess a per patient sensitivity and specificity for tests that aim to detect multiple lesions of which the localization is important. We applied this method to a study designed to evaluate different levels of radiation dose in the detection of polyps by computed tomography colonography (CTC) in patients at risk for colorectal carcinoma (chapter 4). A comparison between different levels of radiation dose based on the per patient sensitivity and specificity could otherwise not have been made. One of the limitations of this method is the somewhat different interpretation of sensitivity and specificity. Furthermore, no diagnostic odds ratio can be calculated from the given sensitivity and specificity. Especially for this application - a comparison between different dose levels - the odds ratios would have been handy.

Synthesizing results from different (diagnostic) studies in systematic reviews can be challenging since many issues have to be considered. One of these is the clinical and methodological heterogeneity between studies. The technique used to synthesize results depends upon the type and the amount of heterogeneity present between studies.\(^4\) When the aim of the meta-analysis is to pool results of tests so to compare these different tests, difficulties arise when different methods have been applied for pooling each test. In chapter 2 a bivariate model is suggested that can address most sources of heterogeneity. An advantage of this approach to meta-analysis is that it gives summary estimates of sensitivity and specificity, the statistics most often used to characterize diagnostic accuracy.

In chapter 5, the bivariate model is applied to pool the results of six different diagnostic tests with the aim see which one is best. When meta-analytic methods will be used for comparing tests it is possible that the results from specific tests are retrieved from comparative studies. In that case a dependency will exist between
the results for tests that have been applied in the same population. In our comparison we did not take this into account. Research is needed to evaluate the magnitude of this dependency and to decide when it is needed to incorporate an extra variable in the model to deal with this dependency.

As mentioned in the general introduction accuracy measures fail to address the added value of a test or the impact a test has on the likelihood of disease. This theme brings us into the field of diagnostic impact or diagnostic thinking efficacy studies. In chapter 6 to 9 subjective probability estimates were used to address such questions. These probabilities express the degree of belief a physician has in a particular diagnosis. They were assigned by physicians on a double logarithmic visual analogue scale.

In chapter 6, subjective probability estimates were used as prior probabilities to study whether probabilities were consistently updated after the test results became available. We were able to use these pre-test probability estimates to identify subsets of patients where the test has added value.

In chapter 7 the probability estimates were used to compare clinical judgment with objective decision tools. Chapter 8 contains a report of a study on how well physicians interpret the ventilation-perfusion lung scan. This was done by calculating the subjective likelihood ratio (LR) from the pre and post-test probability estimates assigned by physicians. Subsequently these subjective LR were compared with the objective LR, obtained form a comparison of the test results with those of the reference standard.

In chapter 9, a transformation of the subjective probabilities was used to specify the amount of information in a test result. This value - entropy - is a measure derived from information theory. It was used to estimate the information content of two tests so to compare them with one another. Entropy allows us to incorporate the information the test provides about the target disease as well as on alternative diagnoses. In addition it is not necessary to define a test result as positive or negative, as was the case in the diagnoses of herniated nucleus pulposus, where there is no unequivocal way to define a negative or positive test result. Whether a test result is labeled as positive will depend on whether the finding on the imaging test matches the pattern of signs and symptoms patients has, as well as on the result of the reference standard. In our study, physicians were only asked to indicate whether their uncertainty had changed after the results of the imaging test had become available. The entropy measure is seldomly used in diagnostic studies. This might be due to difficulties in interpreting this measure. Further research is needed to explore the usefulness of such measures.
**Clinical relevance of this thesis**

Although this thesis focuses on methodology, we feel the results of the studies reported here also have clinical value.

In chapter 4, it is shown that low dose scanners have equal diagnostic accuracy compared to the currently used scanners. At low doses the risk of radiation induced cancer is lower. With these results hopefully manufacturers will be convinced to develop low dose scanners.

In the study reported in chapter 5 we found that most urine-based tumor markers in the detection of bladder cancer lack diagnostic accuracy. In this light cystoscopy should stay the instrument of choice to rule out bladder cancer. In chapter 6 we reported on a study on the added value of the dobutamine stress echocardiography test in patients with chest complaints but with ruled-out acute cardiac disease. This study showed that, although the test has predictive value, the added overall value is low. It is possible the test will perform adequately only in specific subgroups. In chapter 7 clinical judgment was compared with the Ottawa ankle rules clinical decision rules. We illustrated that physicians do as well as objective decision rules in ruling out fractures in patients with sprained ankles, and that they do better in saving unnecessary radiographs. Chapter 8 showed that physicians do well in interpreting ventilation perfusion scans, except for the non-diagnostic test results, where much variation exists in physicians' interpretation. In chapter 9 we found that magnetic resonance imaging holds more information than computed tomography in diagnosing herniated nucleus pulposus, although both tests required additional diagnostic information to decide whether to operate upon a patient.

**Diagnostic Framework**

Ever since Yerushalmy proposed to use the sensitivity and specificity as outcome measures in the evaluation of a tests discriminating performance in 1947 - many efforts have been put into the development of a framework to evaluate new diagnostic technologies. This movement was further stimulated by the notion that medical practice, including the implementation of new technology, needs to be Evidence Based. Such a framework should act analogous to the four phase model used to evaluate new therapies.

So far we have described the results of this thesis in the framework as proposed by Fryback et al. This framework is based on six hierarchical levels a test needs to go through before it should be used in clinical practice. The levels are: technical efficacy (technical properties of the test), diagnostic accuracy efficacy, diagnostic thinking efficacy, therapeutic efficacy (does the test in potential change the
management plan), patient outcome efficacy (does a test increase quality of life) and societal efficacy (cost/benefit evaluations).

None of the proposed frameworks has succeeded in establishing a central role in the process of developing and implementing new diagnostic tests. This is due to many factors, involving both the designs and outcomes used in test research, as well as the nature of diagnostic tests themselves. Many different designs and outcomes are possible depending on the different research questions that can be formulated. Examples are: what is the accuracy of test X, which test, from a series of tests, is best in terms of sensitivity and specificity, does test X have the same accuracy as test Y, what is the best cut-off value of continuous test X, how does test X perform in population K, what does this test add relative to the diagnostic information already available. Detecting a simple hierarchy in these questions is a strenuous task.

Diagnosis in itself is not that straightforward as therapy, where one aims at influencing the course of disease to reduce the patients complaints. The function of diagnostic testing can be described as to reduce the uncertainty of the physician in particular diagnosis. This is often a stepwise process of multiple testing involving many clinical factors. In addition diagnostic tests develop in an enormous rate. Before a test is evaluated well, an even newer version has come up.

A workable framework should be extremely flexible. A suggestion is a framework iterative in character, included the following phases:

Step 1: Development of a test.
Step 2: Understanding the single test in clinical practice.
Step 3: Test within a diagnostic strategy.
Step 4: Implementation, evaluation and evolution.

The research described in this thesis would fit in step 2, except perhaps for chapter 8. This chapter focused more on the evaluation of a specific test, in this case the ventilation-perfusion scan. Based on this research, we found that substantial variation exist in the interpretation of a non-diagnostic test result. These findings could have consequences for the test strategy of diagnosing pulmonary embolism.

We feel that it is important that a workable framework is developed in the near future. This should go together with developing an overview of typology of possible research questions with their designs. Current (lack of) resources does not permit the implementation of poorly evaluated diagnostic tests, which in the end might waist money and even worse might harm patients. Many hospitals already develop guidelines when to implement new diagnostic technology. It would be advantageous if a framework was available where a test should pass through before
it is accepted on the market. Researchers in the field should take the initiative to
develop a useful framework, but governmental bodies (and health insurance
companies) should provide the means to clear the way for medical researchers.

**Clinical Judgment versus Decision Rules**

In chapter 7 we found that physicians performed as well as two clinical prediction
rules in ruling out fractures in patients with acute ankle injury. Later on in chapter
8 we found that physicians did well in overall interpreting the ventilation-perfusion
scans of patients suspected for pulmonary embolism. These findings illustrate that
in specific situations physicians are able to make excellent clinical decisions,
outperforming formal statistic tools.

On the other hand, in clinical practice several initiatives have been deployed
to construct decision aids. These objective tools are thought to lead to less variation
in clinical practice relative to clinical judgment in making decisions. In the end
formal decision aids will yield less medical errors. So the underlying rationale
of the development of decisions aids is that clinical judgment does not suffice in
making proper medical decisions. One other reason for an increase in decision
tools is that the broader availability and knowledge of statistical regression and
neural network techniques stimulated the production of such rules.

Yet it is also shown that such decision aids lack generalizability to other but
related populations compared to the populations where they were derived from. Another disadvantage is that implementation of decision rules so far is expensive
and often unsuccessful. In the future, more research is needed to answer the
question in what situations decision rules might actually be valuable. This should
prevent a waste of time, energy and money in the development, implementation
and evaluation of prediction rules where actual physicians' decision making skills
are more than adequate.

In addition research should be undertaken to unravel the processes of doctoring
itself. To do so, physicians could embrace results from cognitive psychology, the
science that studies clinical judgment. A proper awareness of the products of this
science might accomplish that physicians anticipate better, avoid common pitfalls,
understand repeated mistakes, so they will be more cautious in daily clinical
practice.

In the end an alternate use of decision rules with clinical judgment on evidence
based ground will result in patient benefit as well as in a more intelligent use of
resources.
General Discussion

References