Tubal subfertility and ectopic pregnancy. Evaluating the effectiveness of diagnostic tests

Mol, B.W.J.

Publication date
1999

18. Evaluating the effectiveness of diagnostic tests

Ben W.J. Mol and Patrick M.M. Bossuyt

18.1 Introduction

This thesis reports on two diagnostic problems: the diagnostic work-up of tubal pathology in subfertile couples and the diagnosis of ectopic pregnancy. In the evaluation of the effectiveness of diagnostic tests, clinical research often focuses on the measurement of test performance, i.e., the capacity of a test to discriminate between patients with a particular disease and patients without that particular disease. However, although the discriminative performance of a test is an important determinant of the value of a test, this value depends also on the clinical context of the test, in which the prevalence of disease and available treatments also play a role. If the prevalence of a disease is either very high or very low, additional information to decide whether or not to perform a treatment is probably not required, and the value of the test will be limited. Furthermore, if an effective treatment for a disease that is to be detected is lacking, the value of the diagnostic test will be less than in the situation in which an effective treatment is available.

Many studies evaluating diagnostic tests only report on the discriminative capacity of a test in terms of sensitivity, specificity and likelihood ratios (LR). It may seem obvious that performance of a test with a good sensitivity and specificity on itself does not improve the health status of the patient. The patient will only benefit from adequate treatment, and diagnostic tests can be of value if they distinguish between patients that are likely to benefit from treatment and patients that are likely to be harmed by treatment. The aim of this chapter is to discuss how the effectiveness of a diagnostic test can be determined. The two problems dealt with in the previous parts of this thesis will serve as examples in this discussion.

In § 18.2 the traditional framework for the evaluation of diagnostic tests will be discussed. Parameters for the discriminative capacity of a test, such as sensitivity and specificity, LRs and Receiver Operating Characteristic (ROC) curves, as well as potential pitfalls in the assessment of test performance will be discussed. In § 18.2.1 it will be demonstrated that sensitivity and specificity of a test can vary between patients with different characteristics. § 18.2.2 will deal with the various methods that are available to assemble patients for a diagnostic study, whereas § 18.2.3 will discuss the internal validity of diagnostic studies. § 18.2.4 will assess the relation between reproducibility and accuracy of a diagnostic test. Meta-analysis of diagnostic tests and screening will be discussed in § 18.2.5 and § 18.3, respectively.

Subsequently, § 18.4 will discuss pitfalls in the assessment of the clinical value of diagnostic tests. The value of diagnostic tests can be assessed using decision analysis (§ 18.4.2) or in randomized clinical trials (§ 18.4.3). Finally, in § 18.5 a checklist will be provided that can be used in the critical appraisal of articles reporting on diagnostic tests.

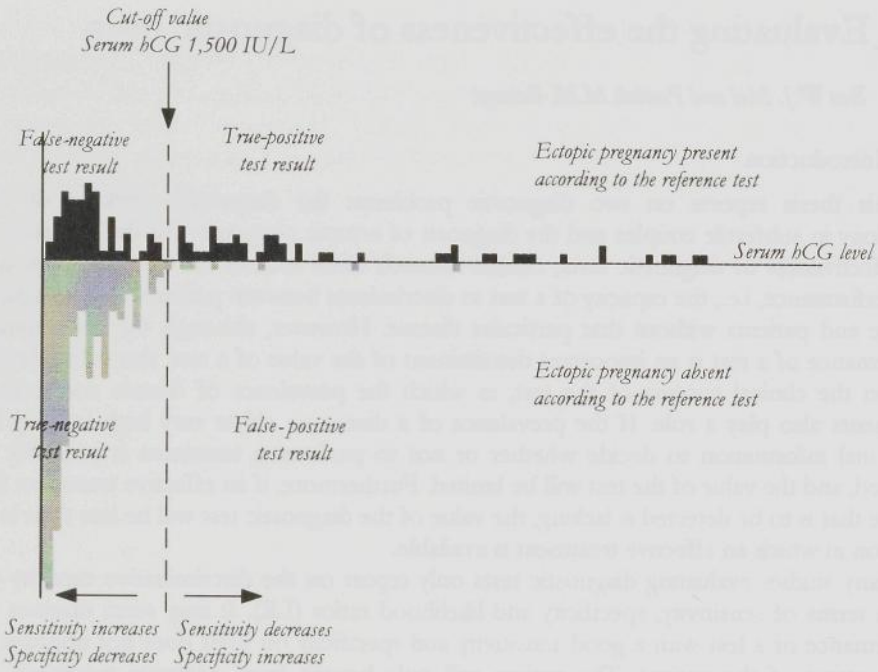

Figure 1: Distribution of serum hCG level in patients with and patients without an ectopic pregnancy, and inconlusive findings at transvaginal sonography.

18.2 Expressing accuracy of diagnostic tests

Since in most situations a diagnostic test does not perfectly discriminate between people that have the disease and people that have not, it is important to know how accurate the test informs the clinician about the disease status of a patient. The accuracy of a diagnostic test is usually expressed in terms of sensitivity and specificity, ROC-curves or LRs. These measures are obtained from patient series in which both the test under study and a second test are performed. The latter is often referred to as 'gold standard' test, since it is assumed to provide information on the presence of disease with 100% accuracy. However, this second test is often only an approximation of the true disease state, and use of the term 'gold standard' is not always justified. It should therefore preferably be replaced by the term 'reference test'.

The sensitivity of the test under study is defined as the number of people with a positive result of the test under study and a positive reference test result, divided by the number of people with a positive reference test result. The specificity of the test under study is defined as the number of people with a negative result of the test under study and a negative reference test result, divided by the number of people with a negative reference test result.

In the medical literature, 'positive' and 'negative' test results are in this context always referring to the presence and absence of disease, respectively. Consequently, a clinician labeling a test result as 'positive' refers to something that is usually negative for the patient. A 'negative' test result on the other hand refers to something that is usually 'positive' for the patient. Thus, from a patients' perspective it would make more sense to swap the use of these terms. However, since the description of test performance historically has focused on the detection of disease the term 'positive' is used to refer to presence of disease, and 'negative' is used to refer to absence of disease. This can sometimes be contra-intuitive. For instance, in the diagnosis of ectopic pregnancy, absence of an intra-uterine pregnancy at sonography refers to the presence of ectopic pregnancy and is referred to as positive test result, whereas presence of an intra-uterine pregnancy at sonography refers to the absence of ectopic pregnancy and is referred to as negative test result

A limitation of the use of sensitivity and specificity is that they require dichotomization of the results of diagnostic tests into 'positive' and 'negative'. Many diagnostic tests generate more detailed information than 'disease present' or 'disease absent'. This information can be used to estimate the probability that a patient has the disease as exactly as possible. Figure 1 shows the distribution of serum human chorionic gonadotrophin (hCG) in patients with and patients without ectopic pregnancy, in whom transvaginal sonography did not reveal a definite diagnosis. As can be seen in Figure 1, the serum hCG concentration can be expressed on a continuous scale. The mean serum hCG concentration is higher in women with an ectopic pregnancy than in women without an ectopic pregnancy. By setting a 'cutoff' value we can define the results of the serum hCG measurement as either positive or negative. Comparison of this test result with the results of a reference test enables a classification of test results as true-positive, false-positive, true-negative and false-negative, thereby allowing the calculation of sensitivity and specificity for this specific 'cutoff' value. Figure 1 shows that a cutoff value of 1,500 IU/L results in only a few false-positive diagnoses, whereas a considerable part of the women with ectopic pregnancy is correctly identified.

A shift in cutoff value can alter the sensitivity and specificity of serum hCG measurement (Figure 2). A shift in cutoff value towards higher serum hCG levels increases the specificity but decreases the sensitivity. Vice versa, a shift in cutoff value towards lower serum hCG levels decreases the specificity but increases the sensitivity of a test. It is this phenomenon on which the principle of the ROC-curve is founded. By graphing the sensitivity and specificity for each possible cutoff value in ROC-space, a visual impression can be obtained of the discriminative performance of a diagnostic test. In such a ROC-space the sensitivity (or true-positive rate) is plotted against '1 – specificity' (or false-positive rate). The upper left hand corner of such a ROC-space, in which sensitivity and specificity are both 100%, represents a test result with perfect discriminatory performance, whereas the line sensitivity = '1 – specificity' (the dashed line in Figure 2) represents a test without any discriminatory performance. The more the dis-

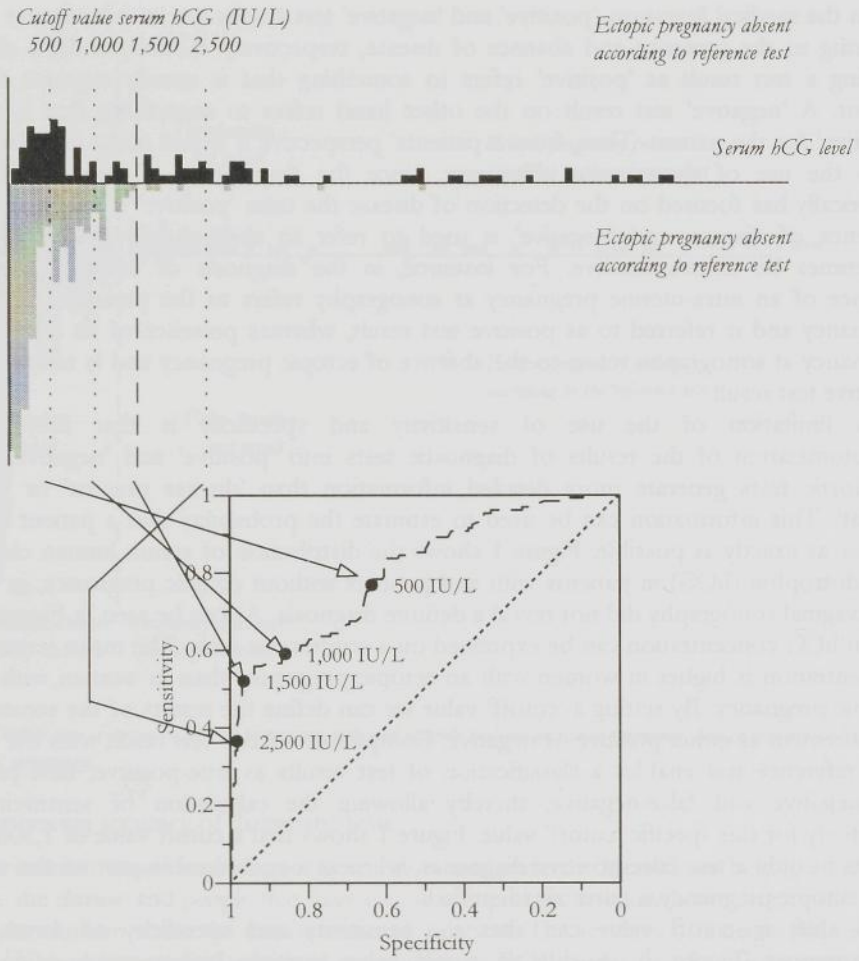

Figure 2: The relation between the distribution of a test result in diseased and non-diseased persons (here patients with and patients without an ectopic pregnancy) and a corresponding ROC-curve that expresses the performance of the test.

tributions of the marker differ between diseased and non-diseased, the better the performance of a diagnostic test becomes, and the larger the area under the ROC-curve will be. Use of a single cutoff value will always result in loss of information, which might be of importance in the diagnostic process.

Figure 2 shows that with higher serum hCG concentrations, the probability that a woman has an ectopic pregnancy increases. For example, in case the serum hCG concentration is 2,500 IU/L, the probability that a woman has an ectopic pregnancy is higher than in case the serum hCG concentration is 1,600 IU/L. However, if 1,500 IU/L

Table 1: Effect of variation of the prevalence of ectopic pregnancy among women with different characteristics on predictive values of transvaginal sonography (TVS) of the adnexal region. The sensitivity and specificity of this test are 57% and 96%, respectively. In the left two-by-two table the situation is shown for a prevalence of ectopic pregnancy of 6%, corresponding with the prevalence among symptom-free women at increased risk for ectopic pregnancy. The positive predictive value (PPV) equals 47% (34/72), whereas the negative predictive value (NPV) equals 2.8% (26/928). In the right two-by-two table the situation is shown for a prevalence of 40%, corresponding with women with abdominal pain in early pregnancy. The PPV changes to 90% (228/252), whereas the NPV changes to 23% (172/748).

TVS	Ectopic Pregnancy	No ectopic pregnancy
Test positive	34	38
Test negative	26	902

Prevalence 6%; PPV = 47%; NPV = 2.8%

TVS	Ectopic pregnancy	No ectopic Pregnancy
Test positive	228	24
Test negative	172	576

Prevalence 40%; PPV = 90%; NPV = 23%

is used as cutoff, both serum hCG concentrations of 1,600 IU/L and 2,500 IU/L are interpreted in a similar way.

A method to overcome the loss of information by dichotomizing a test result is the use of LRs. A LR of a particular test result is defined as the ratio of the relative frequency of that particular test result in patients in which the result of the reference test is positive and the relative frequency of that particular test result in patients in which the result of the reference test is negative. A major advantage of the use of LRs is that they can be applied easily in daily clinical practice. By integrating the pre-test probability on disease and the LR of a certain test result in Bayes Theorem, the probability that a patient has the disease conditional on the presence of a certain test result can be calculated. In *chapter 10* and *11* LRs were calculated for different categories of serum hCG concentrations at initial and repeat measurement, respectively.

The LR of a particular test result is mathematically equal to the ratio of sensitivity and '1-specificity'. Since in a ROC-curve the sensitivity is plotted against '1-specificity', the LR of a test result equals the slope of the ROC-curve. Thus, apart from the area under the curve, the ROC-curve also provides information on the diagnostic performance of a test by its slope. The LR of a range of test results can be derived from the slope of the line connecting the two points that correspond with the two most extreme test results of that range.

Since sensitivity and specificity express the relative frequency of a positive test result in patients with a positive reference test, and the relative frequency of a negative test result in patients with a negative reference test, respectively, these parameters are by definition independent of the prevalence of disease. Thus, if sensitivity and specificity of a particular test are determined in a certain study, they can be used in other populations with a different prevalence of disease for the calculation of predictive values. This is important, since the clinician that is using the test in clinical practice will be mostly interested in these

predictive values. The predictive value of a test result is defined as the probability of a positive reference test conditional on the test result. The positive predictive value represents the probability of a positive reference test, in case the test result is positive. Similar, the negative predictive value represents the probability of a positive reference test in case the test result is negative.

Table 1 shows the effect of variation of the prevalence of disease on the predictive value of a test. In *chapter 15* the prevalence of ectopic pregnancy among symptom-free women at increased risk for this disease was found to be almost 6%. Presence of a gestational sac or an ectopic mass at transvaginal sonography had a sensitivity of 57% and a specificity of 96%. A positive result of transvaginal sonography, i.e., presence of a gestational sac or an ectopic mass at transvaginal sonography, performed in a population with a prevalence of the disease of 6% implicates a positive predictive value of 47%. A negative result of a test with such sensitivity and specificity implicates a negative predictive value of 2.8%. The same test performed in women with abdominal pain, in which a prevalence of ectopic pregnancy of 40% was observed, results in positive and negative predictive values of 90% and 23%, respectively. Such differences in predictive values are very likely to have clinical consequences.

When calculating predictive values for patients in a population with a certain prevalence, this prevalence is likely to be related to the patient characteristics in that specific population. In *chapter 12*, LRs of results of transvaginal sonography and serum hCG measurement were used in patients with different combinations of clinical symptoms, for which different prevalences of ectopic pregnancy apply. This resulted in an improved performance as compared to an algorithm using fixed cutoff values, especially for symptom-free women, in whom the pre-test probability for ectopic pregnancy is rather low.

18.2.1 Effect of patient characteristics on sensitivity and specificity

Since there appears to be a clear association between clinical symptoms and prevalence, the crucial underlying assumption when applying sensitivity and specificity on patients with different prevalences is that these indices remain constant for patients with different clinical characteristics.[1,2]

In *chapter 10*, it was therefore evaluated if the performance of serum hCG measurement in patients with inconclusive findings at sonography depended on patient characteristics. Whereas this performance seemed independent from the presence or absence of abdominal pain or vaginal bleeding, it was found to depend on the presence of an adnexal mass or fluid in the Pouch of Douglas at transvaginal sonography.

Although the relation between sensitivity, specificity, prevalence and predictive values is often demonstrated in textbooks, the assumption that the indices sensitivity and specificity remain constant for different types of patients is only rarely evaluated in diagnostic studies. Apart from the example of serum hCG measurement in this thesis, the constancy assumption for sensitivity and specificity has also found to be erroneous in other clinical situations where it has been checked, the best documented example being exercise tests in the diagnosis of ischaemic heart disease.[3-6]

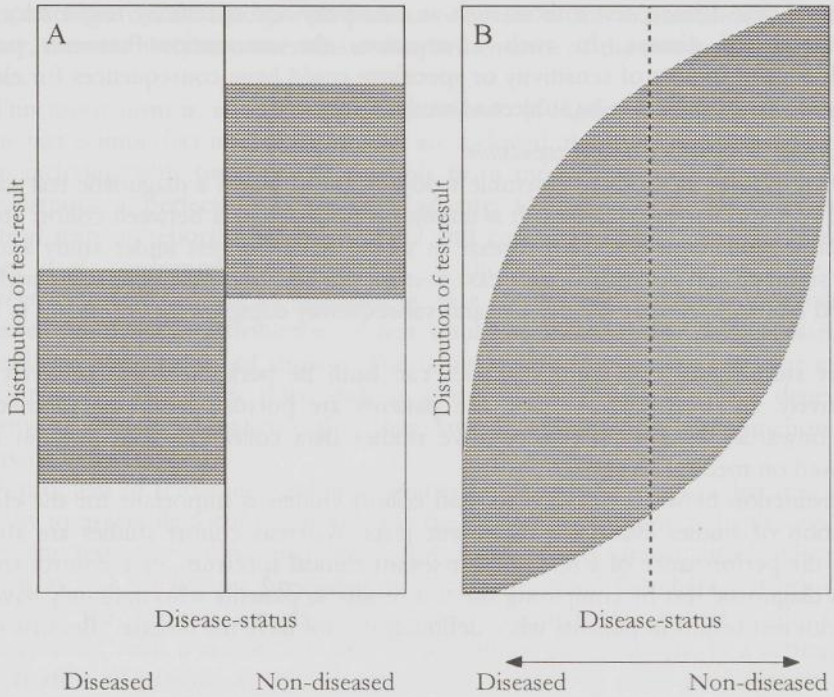

Figure 3: Explanation for the variation of sensitivity and specificity with patient characteristics. Figure 3A shows the situation in which the distribution of test results differ among patients with the disease and patients without the disease, but not within the group of patients with or without the disease. Figure 3B shows the situation in which the distribution of test results varies with the severity of disease. Here, sensitivity or specificity can vary between patients with different characteristics.

A theoretical explanation for the variation in sensitivity and specificity between different types of patients is illustrated in Figure 3. Figure 3A shows the situation in which the results of a test differ between patients with the disease and patients without the disease, but not within the group of patients with the disease and within the group of patients without the disease, respectively. Consequently, sensitivity and specificity remain constant, even in situations in which the prevalence of disease in the population in which the test is used differs from the prevalence in the population in which the test was evaluated.

Figure 3B shows a different situation. The results of the test under study shown in this figure depend on the severity of the disease. In other words, within the group of patients with the disease the test results become more extreme in patients in whom the disease is more severe. If the results of a test in patients with the disease do depend on the severity of disease, the sensitivity of a test in patients with a mild form of the disease might differ from the sensitivity of a test in patients with a severe form of the disease. The same appears to be true for the specificity. Thus, if the severity of disease is related to the

prevalence of that disease, test indices such as sensitivity and specificity might vary with the prevalence of disease. In such a situation, the association between patient characteristics and indices of sensitivity or specificity could have consequences for clinical practice, and should therefore be subject of study.

18.2.2 Methods to assemble a patient population

There are several methods to assemble a population in which a diagnostic test can be compared with a reference standard. It is important to distinguish between cohort studies that assemble patients at risk for a disease in whom both the test under study and the reference standard are performed, and case-control studies, that assemble cases with the disease and controls without the disease, and subsequently compare the test result in both groups.[1]

Cohort studies and case-control studies can both be performed prospectively and retrospectively. In prospective studies, the patients are pursued from the moment of inclusion onwards, whereas in retrospective studies data collection goes back in time, mostly based on medical charts.

The distinction between case-control and cohort studies is important for the clinical interpretation of studies evaluating diagnostic tests. Whereas cohort studies are able to report on the performance of a test in the relevant clinical spectrum, case-control studies evaluate a diagnostic test by comparing the test results in patients who definitely have the disease with test results in patients who definitely do not have the disease. Because case-

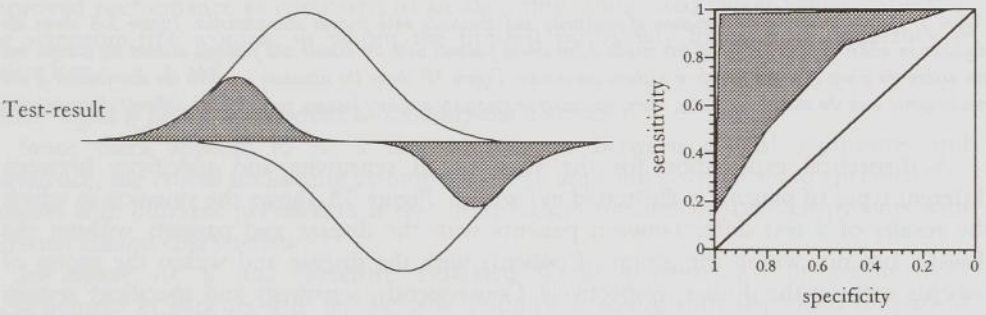

Figure 4: Effect of methods to assemble a patient population on the estimated performance of a diagnostic test, as expressed with a ROC-curve. The left figure shows the distribution of a test result in both case-control and cohort studies, and the right figure shows the corresponding ROC-curves. The gray area represents a distribution of a test-result in a population assembled in a case-control study. Since the test results are obtained in patients rather at the extremes of the clinical spectrum, the ROC-curve demonstrates excellent diagnostic performance. In contrast, the white area represents test results of patients assembled in a cohort study, and the corresponding ROC-curve in these patients demonstrates limited diagnostic performance.

control studies usually incorporate persons who are at the extremes of the spectrum of disease, they are likely to overestimate the performance of the test as compared to cohort studies.

This mechanism is, in an extreme form, shown in Figure 4. Whereas the performance of the test is imperfect in case all patients are included, the ROC-curve that is constructed when including only patients and controls from the extremes of the disease spectrum demonstrates a perfect discriminative capacity. Comparison of the performance of identical tests as reported by case-control and cohort studies, as was recently done by Lijmer *et al.*, confirmed this hypothesis.[7] It should be emphasized that the difference between test performance as reported by case-control and cohort studies can only be observed in case the distribution of test results in diseased and non-diseased persons depends on the severity of disease. The latter implies that parameters that express the performance of diagnostic tests, such as sensitivity, specificity and LRs, depend on the severity of disease of patients that they are performed in, a phenomenon that was discussed in § 18.2.1.

A problem of retrospective cohort studies on diagnostic issues is that it is sometimes difficult to assemble a cohort at risk for the disease in retrospect. In the evaluation of diagnostic tests for ectopic pregnancy, for example, it is almost impossible to identify women at risk for ectopic pregnancy in retrospect, since the characteristics of women suspected for ectopic pregnancy are not systematically collected in a registry. Consequently, such a cohort can only be assembled prospectively, in a setting in which identification of patients at risk for the disease is done by study investigators who have time to check whether patients fulfill the inclusion criteria. In contrast, the cohort of subfertile women in whom the performance of hysterosalpingography (HSG) was studied could be assembled in retrospect, since the medical administration systematically registers patients in whom HSG has been performed (*chapter 5*). In contrast, case-control studies can easily be performed retrospectively with respect to inclusion of patients. Consequently, such studies can be completed in a shorter period of time at relatively lower costs.

A second problem in the performance of retrospective studies, be it either cohort or case-control studies, is that sometimes the results of the test under study are difficult to retrace. For example, the evaluation of the capacity of HSG to predict fertility outcome in *chapter 5* was hampered by the fact that intra-uterine abnormalities observed at HSG were not registered in the medical files in a systematic way. Another example is the evaluation of transvaginal sonography in the diagnosis of ectopic pregnancy, a study that is virtually impossible to perform retrospectively, since results of sonography are not documented systematically.

The extent to which case-control studies overestimate the performance of a diagnostic test depends on the nature of the disease that the test is trying to detect. In case the distribution of test results between diseased persons included in a case-control study (the cases), and diseased persons included in a cohort study is different, the sensitivity reported by case-control studies might differ from the sensitivity reported in cohort studies. Similarly, if the distribution of test results differs between non-diseased patients included

in a case-control study (the controls) and non-diseased persons included in a cohort study, the reported specificity might differ. For example, when assessing the performance of CA-125 measurement in the diagnosis of endometriosis, studies comparing CA-125 levels in cases with severe endometriosis and controls who undergo laparoscopy for sterilization, do not take into account that endometriosis has various presentations. Understandably, a meta-analysis assessing the capacity of CA-125 measurement to detect endometriosis found large differences in the performance of CA-125 measurement reported in case-control and cohort studies.[8]

18.2.3 Internal validity

Several forms of bias can occur in studies evaluating diagnostic tests. *Verification bias* or *ascertainment bias* occurs when the decision to perform a reference test is at least partially dependent on the result of the test under study.[1,9] This is important in the assessment of a diagnostic test when two different reference tests are used to verify a diagnosis. In the study described in *chapter 10*, for example, the decision to perform laparoscopy or to manage a patient expectantly depended at least partially on the serum hCG level. An ectopic pregnancy was therefore more likely to be established in patients with a serum hCG level ≥ 1,500 IU/L than in patients with a serum hCG level < 1,500 IU/L. Since expectant management can be considered as the ideal reference strategy in determining which ectopic pregnancy requires treatment and which ectopic pregnancy does not require treatment, performance of laparoscopy in patients with a serum hCG ≥ 1,500 IU/L leads to overestimation of sensitivity and specificity of serum hCG measurement. Studies reporting on the outcome of expectant management in patients with suspected ectopic pregnancy are needed to reveal this issue, if only they report on the relation between serum hCG concentration and failure of expectant management.

Selection bias occurs when inclusion of a patient in the study depends (at least partially) on the result of the test under study. Table 2 shows the mechanism by which selection bias can affect the estimates of sensitivity and specificity. In absence of selection bias, all patients that meet the inclusion criteria of the study are incorporated in the final two-by-two table, independent of the result of the test under study. Once the test results are compared with the results of the reference strategy (left two-by-two table), sensitivity and specificity can be calculated. If in case of a negative test result a patient is less likely to be included in the study than in case of a positive test result, the fraction of patients with a negative test result is relatively low as compared to the fraction of patients with a positive test result. As can be seen in the right part of Table 2, this will result in an overestimation of the sensitivity and an underestimation of the specificity of the test under study. The opposite would have been the case if patients with a positive test result had been less likely to be included as compared to patients with a negative test result. If the true ratio of patients with a positive test result and patients with a negative test result in the original population under study is known, the true estimates of sensitivity and specificity can be derived, as is shown in the right part of Table 2.[9] However, if this ratio of test results is unknown, the true sensitivity and specificity can not be derived.

	Disease Present	Disease absent	
Test positive	A	C	A + C
Test negative	B	D	B + D

$Sensitivity_{true} = A / (A + B)$
$Specificity_{true} = D / (C + D)$

	Disease present	Disease absent	
Test positive	A	C	A + C
Test negative	B – b'	D – d'	(B – b') + (D – d')

$Sensitivity_{obs} = A / (A + (B - b'))$
$Specificity_{obs} = (D - d') / (C + (D - d'))$
$Sensitivity_{true} = A / (A + (B - b')*\gamma^{-1})$
$Specificity_{true} = (D - d')*\gamma^{-1} / (C + (D - d')*\gamma^{-1})$

Table 2: Effect of selection bias on the sensitivity and specificity. Left the situation without selection bias. All patients that undergo the test under study are included, and sensitivity and specificity can be calculated. Right the situation after selection bias. A proportion b' of the false-negatives and a proportion d' of the true negatives do not undergo the reference test since they had a negative result of the test under study. If the percentage of patients γ (= (b' + d')/(B + D) that did not undergo the reference test is known, the true sensitivity and specificity can be calculated.

Selection bias can also occur if patients are selected on the basis of a test result that is (partially) related to a result of the test under study. For example, if a study wants to evaluate the diagnostic accuracy of HSG by comparing it with laparoscopy, and if the *Chlamydia* antibody titer (CAT) was used in the decision to include a patient in the study, the association between HSG and CAT might lead to biased estimates of sensitivity and specificity.

Incorporation bias occurs when the test under study is used as part of the reference strategy. Understandably, this will lead to overestimation of both sensitivity and specificity of the test. In order to avoid incorporation bias, sonography at a gestational age of 12 weeks was performed in all patients with suspected ectopic pregnancy in whom sonography previously had visualized an intra-uterine pregnancy *(chapter 10 and 11)*.

Lack of blinding can cause bias in two ways. First, if the person interpreting the reference test is aware of the result of the test under study, this knowledge is likely to influence the interpretation of the reference test. This type of bias is called '*diagnostic review bias*'. Second, the person who is performing or interpreting the test under study might already have knowledge of the result of the reference test. This type of bias is called '*test review bias*'.

Finally, bias can occur when patients with inconclusive test results or patients in whom the test fails are excluded from the final analysis. This will in most cases lead to an overestimation of sensitivity and specificity of a test.

18.2.4 Explanations for false-positive and false-negative diagnoses

The vast majority of tests used in clinical medicine is imperfect. To understand the cause of such diagnostic imperfection, it is important to distinguish factors that determine

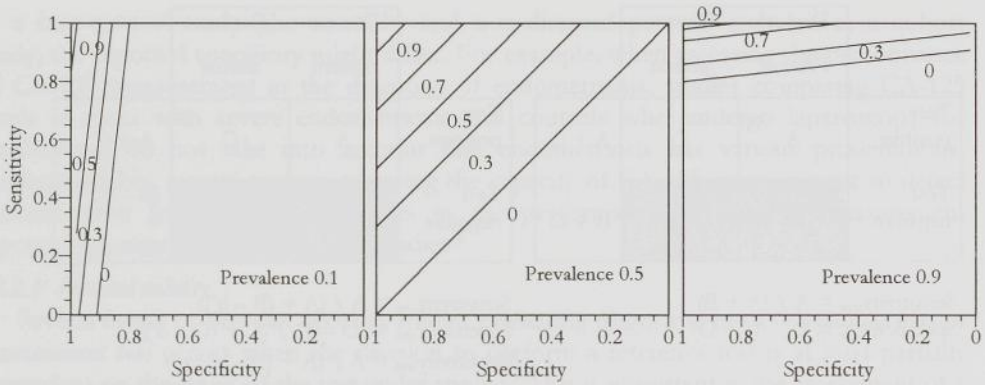

Figure 5: Relation between sensitivity, specificity and reproducibility of a test (expressed by kappa-values) for prevalences of 0.1, 0.5 and 0.9, respectively. The corresponding kappa-values are plotted near the plots of sensitivity and specificity.

the performance of a test. First, the disorder to be detected must have pathophysiological or anatomic features, which allow its detection by the test. Second, the test must have the ability to discriminate between subjects with and subjects without that feature. Third, the physician must interpret the results of the test correctly. In case the observer classifies a test result incorrectly as abnormal or incorrectly as normal, this will affect the specificity and sensitivity of the test, respectively. It is therefore inevitable that if the reproducibility of a test is not perfect, the accuracy of the test can not be perfect either. Lack of reproducibility of a test within one observer means that if this observer reads out several test results, there will always be some false-positive or false-negative results. Lack of inter-observer reproducibility of a test indicates that if the test is used by multiple clinicians, there will always be false-positive or false-negative results, even if there may be one observer who reads out the test result without mistakes.

Several authors have assessed the relation between reproducibility and accuracy mathematically.[2 10] If the reproducibility of a test is expressed as a kappa-value it is possible to calculate the best possible combinations of sensitivity and specificity for that particular test. The kappa-value (κ), that expresses agreement beyond chance, can be calculated from:

$$(I)\quad \kappa = (\text{observed agreement} - \text{expected agreement}) / (1 - \text{expected agreement})$$

in which the observed agreement equals

$$(II)\quad \text{sensitivity} * P_{Dis} + \text{specificity} * (1 - P_{Dis})$$

and the expected agreement equals

$$(III)\quad P_{Dis} * P_{test\_pos} + (1 - P_{Dis}) * P_{test\_neg}$$

Integrating (II) and (III) into (I) shows that the sensitivity equals

$$(IV) \quad (\kappa - (\kappa + 1)*(P_{Dis} * P_{test\_pos} + (1 - P_{Dis}) * P_{test\_neg}) - specificity * (1 - P_{Dis}))/ P_{Dis}$$

When the prevalence of disease P_{Dis} is assumed to be equal to the probability of a positive test result P_{test_pos}, the most optimal combinations of sensitivity and specificity can be calculated when the prevalence of disease and the reproducibility of a test as expressed by a kappa-value are known. In Figure 5 the most optimal combinations of sensitivity and specificity are plotted for different kappa values and prevalences of 0.1, 0.5 and 0.9, respectively. One must realize that the plotted combinations of sensitivity and specificity are the most optimal combinations conditional on a known prevalence and reproducibility of a test. As mentioned, there are other factors that can affect the accuracy of a test, in addition to reproducibility. In *chapter 2*, for example, the reproducibility of the diagnosis of tubal occlusion at HSG was found to be almost perfect. However, the sensitivity of the test was still found to be 65% for a specificity of 83%. Apparently, other factors also affected the accuracy of HSG in the diagnosis of tubal occlusion.

The reproducibility of continuous tests is preferably expressed as an intra-class correlation coefficient. The relation between reproducibility and accuracy can also be described using intra-class correlation coefficients.[11] The relation between reproducibility and accuracy is particular useful in situations in which the reference test is expensive, or in which there is no good reference test at all. In case the reproducibility a test appears to be suboptimal, it might become clear that the test is not suitable for clinical practice. In *chapter 2*, we found the reproducibility of adhesions to be moderate to substantial, which implies that its accuracy can never be perfect. This finding was confirmed in the meta-analysis of *chapter 3*.

18.2.5 Meta-analysis of diagnostic tests

The performance of a diagnostic test will often have been addressed in multiple studies, performed in different hospitals, in different settings and at different moments in time. Whereas the accuracy of the estimations of test performance in each of the individual studies might be limited, combined analysis of the results of these studies can increase the accuracy of the assessment of the test under study. Meta-analysis is a tool that can be used to combine results of individual studies. Apart from the increased accuracy, it also facilitates exploration of the impact of covariates on the performance of the test under study. Covariates might be issues of study design, characteristics of the population in which the test is evaluated or characteristics of the test itself.

In the evaluation of effectiveness of therapy meta-analytic methods have been used to summarize results of multiple studies on many occasions.[12] Although controversy remains on the value of such tools as compared to large, multi-center randomized clinical trials, meta-analysis is a powerful tool in absence of such large multi-center studies.[13] The methodology of meta-analysis of diagnostic tests is still under development. Recently, efforts have been made to evaluate meta-analytic methods for the assessment of diagnostic test performance as reported in multiple studies.[14-16]

Apart from heterogeneity due to differences in study design, patient population or technical aspects, heterogeneity in studies reporting on the performance of a diagnostic test might be due to differences in cutoff levels that are used in different studies. If this were to be the case, higher sensitivity would be accompanied by a lower sensitivity and vice versa. Such a hypothesis can be evaluated by visual inspection of the results of the different studies in a ROC-sheet, or, more formally, by calculation of a Spearman correlation coefficient. If such a relation between sensitivity and specificity is found, the accuracy of the test under study can be expressed in a summary ROC-curve, as was done in the meta-analyses on the performance of HSG and CAT in *chapter 3* and *4*.[15]

Taking into account the two aims of meta-analysis for diagnostic tests, i.e., increase of the accuracy of the estimation of the performance of the test under study and exploration of the impact of covariates on the performance of the test under study, exploration of co-variates should be the first step in meta-analysis of diagnostic tests. This can be done by comparing the diagnostic odds ratios (OR) of studies with a certain characteristic and studies without that characteristic. Logistic regression analysis is a powerful tool in the comparison of such diagnostic ORs. If the diagnostic ORs of studies with and studies without a certain characteristic were to be different, further meta-analysis should be performed in subgroups, or should be limited to the subgroup of studies that is the most relevant from a clinical or methodological perspective.

As an alternative, the individual studies can be scored according to predefined criteria that address the methodology of the study, for example issues of patient sampling, data collection, blinding or presence of bias. For each study, a score can than be established, that represents the methodological quality of such a trial.[17] Such quality scores can than be used as weighting factors in the meta-analysis.

An unresolved problem in the scoring for methodological quality, however, is the quantification of different criteria. Whereas in studies on one particular diagnostic test a certain criterion might be of extreme importance, this same criterion might of less importance in studies on another diagnostic test. Blinding, for example, is an important issue in studies comparing the accuracy of HSG or CAT, using laparoscopy as the reference standard, as was the case in the meta-analysis performed in *chapter 3 and 4*. In contrast, in studies relating HSG findings or laparoscopy to the occurrence of pregnancy or live-birth *(chapter 5 and 6)* blinding might be of less importance.

Taking into account these unresolved problems in the quantification of the methodological quality of studies, an approach in which studies incorporated in a meta-analysis are scored for methodological issues in a qualitative way seems to be preferable. The importance of these criteria on the outcome of a meta-analysis of a particular test can subsequently be evaluated by means of logistic regression analysis, as was done in the meta-analysis of CAT (*chapter 4*), as well as in other meta-analyses.[8 18]

Exploration of co-variates with logistic regression analysis being the first step in meta-analysis of diagnostic tests, the second step is to assess if calculation of point estimates of sensitivity and specificity is meaningful. This is the case if homogeneity cannot be rejected. In absence of homogeneity, the possibility of estimating a summary ROC-curve should be

explored. If studies with a higher sensitivity appear to have lower specificity and vice versa, calculation of a summary ROC-curve is justified.

A major problem of the meta-analytic method for diagnostic tests is that the estimated summary ROC-curves do not provide cutoff values, thereby limiting the implementation of the results of meta-analysis for clinical practice. Whereas current meta-analytic methods allow comparison of the performance of several diagnostic tests, it is not possible to link cutoff values to the estimated summary ROC-curve. The increasing facilities for electronic data management and electronic communication might reduce the need for such meta-analytic methods in the near future, since opportunities will be taken for aggregate analysis of individual patient data collected in different studies.

18.3 Screening versus diagnosis

Screening is defined as systematic detection of disease before the onset of symptoms by means of a relatively simple test.[19] The essential difference between diagnosis and screening is that in case of diagnosis the patient contacts the health care provider, usually after the onset of complaints, whereas in case of screening the health care provider contacts the person who is at risk for having a disease or a preliminary stage of it.[20] This difference is important since it indicates that in case of diagnosis the patient is already worried about her health condition, whereas in case of screening the patient is unaware of a possible disease. Thus, when considering to offer screening one should take into account the additional anxiety that is induced by informing the woman that she may have a disease. The disutility that is caused by this additional anxiety should be compensated for by the expected health benefits of screening.

The difference between diagnosis and screening is illustrated when the situation of a woman that presents with vaginal bleeding during first trimester pregnancy is compared with the situation of a women during first trimester pregnancy without complaints, who has had tubal surgery in the past. The woman with vaginal bleeding has contacted her physician because she is worried about her pregnancy, and her physician should help her as good as possible. In contrast, the woman with a history of tubal surgery has been previously informed that she is at increased risk for having an ectopic pregnancy once she has conceived. The physician that informs the woman about this increased risk can only do so if the expected benefit from screening women in this situation outweighs the expected harm caused by the anxiety or a false-positive diagnosis. In this example, the concern of the women with vaginal bleeding is an extra argument to perform transvaginal sonography. In contrast, the woman with a history of tubal surgery should first be informed about the possibility that her pregnancy might be ectopically nidated.

There is considerable debate about the valuation of false-positive test results in screening. Since a person who undergoes screening has no complaints and is usually healthy before the start of screening, some argue that harm done to such a person should be valued worse as compared to the same amount of harm done to a person who is presenting with complaints, and for that reason is undergoing diagnostic tests.[20] For example, a false-positive diagnosis leading to laparoscopy in a woman suspected of

ectopic pregnancy because of vaginal bleeding might be valued less severe as compared to a false-positive diagnosis leading to laparoscopy in a woman suspected of ectopic pregnancy because a history of tubal surgery. Such differences in valuations will always be arbitrary. In the decision analysis in *chapter 16*, in which screening for ectopic pregnancy in symptom-free women at increased risk was compared with watchful waiting we valued a false-positive diagnosis in a symptom free woman in the same way as a false-positive diagnosis in a woman with vaginal bleeding. In spite of this assumption, the possible benefits of screening did not unequivocally outweigh its possible detriments. Thus, in case the negative impact of a false-positive diagnosis of screening would be valued stronger, the merits of screening for ectopic pregnancy would even become more questionable.

18.4 Assessment of the clinical value of diagnostic tests

After having assessed the performance of a diagnostic test, taking into account possible sources of bias, the subsequent question is whether a diagnostic test is valuable for clinical practice. If, for example, serum hCG measurement in the diagnosis of ectopic pregnancy has an area under the ROC-curve of 0.81, does this mean that this test is valuable for clinical practice? And does the fact that in patients without any findings at transvaginal sonography this area is 0.74, whereas it is 0.85 in patients with free fluid or an adnexal mass, have an impact on the clinical value of this test? In the evaluation for tubal pathology in subfertile patients, does the fact that HSG has a sensitivity of 65% and a specificity of 83% imply that we should use this test in clinical practice?

Some authors have provided scoring systems in order to assess the value of a diagnostic test by its LR. [21] For positive test results, a LR < 2 would indicate that the test is useless, a LR between 2 and 5 would implicate that the test has some value, whereas higher LRs would increase the value of the test. Unfortunately, such an approach does not take into account the clinical context of the test under study. As was shown in § 18.2, predictive values, which are of main interest for the clinician, do not only depend on the performance of a test, but also on the prevalence of disease. Presuming constant indices for sensitivity and specificity, the positive and the negative predictive values decrease when the prevalence of disease decreases.

Apart from the prevalence of disease, the value of a diagnostic test also depends on the availability of an effective treatment in patients with the disease that the test under study is trying to detect. If the side-effects of a particular treatment are mild or if its costs are low, the consequences of a false-positive diagnosis, which would result in treatment of a person without the disease are limited. If, however, a particular treatment generates severe side-effects or if its costs are high, the consequences of a false-positive diagnosis are far more serious. Similarly, the consequences of a false-negative diagnosis are limited in case delay of the required treatment in a patient has a small impact on the outcome for the patient, whereas the consequences of a false-negative diagnosis are far more serious when delay of effective treatment affects the prognosis for the patient. Thus, the valuation of consequences of false-positive and false-negative test results has an impact on the value of a diagnostic test. Apart from the prevalence of disease and the performance of the test,

the therapeutic context, i.e., benefits and harm of - incorrect - treatment and non-treatment, are of importance in the assessment of the value of the test. Furthermore, the costs and harm of the test itself affect the value of the test.

The fact that a diagnostic test can only become of value for the patient if it improves her health status, has consequences for test evaluation. Considering the therapeutic context, a diagnostic test reduces uncertainty on the disease status, thereby increasing the foundation for the decision to provide or withhold treatment. The aim of diagnostic testing is then to increase the probability of presence of disease in a subgroup of patients in such a way that the expected potential benefits of treatment outweigh the expected potential harm of treatment, or to decrease the probability of presence of disease in such a way that the expected potential benefit of non-treatment outweighs the expected potential harm of non-treatment.

Apart from its impact on therapeutic decisions, the information provided by diagnostic tests can also be of direct value for the patient. Patients may want to be informed about the cause of their disease or about their prognosis. For example, pregnant women with a normal pregnancy are willing to pay for information contained at ultrasound, and almost 50% of the value of the ultrasound was pertained to uses outside the realm of medical decisions.[22] The assessment of this 'informative' value of a diagnostic test makes other demands on evaluation of diagnostic tests, which are beyond the scope of this thesis.[23] Generally, it is thought that better information on disease status and prognosis are valuable to the patient. The work-up for subfertility and the diagnosis of ectopic pregnancy have never been explicitly studied from this perspective. One could hypothesize that the work-up for subfertility provides information on the causes for sub- or infertility, which can be of importance for the couple. In the diagnosis of ectopic pregnancy it is likely that women benefit from ultrasound, even if the ultrasound findings do not affect the management of such patients. This is especially the case for patients with a history of ectopic pregnancy or subfertility. Some empirical studies that have assessed the value of information itself, have indicated that an increase of information does not always for the patient. A recent study among women referred for colposcopy after an abnormal PAP-smear showed that women receiving simple information are less anxious than women either receiving no information or more complex information.[24]

If we do not take into account this 'informative' value of a diagnostic test, but limit the use of information to its impact on the treatment strategy, the value of a test can be assessed in two possible types of studies: decision analysis or randomized clinical trials. These frameworks have in common that in both of them the performance of a diagnostic test under study can be related to health states that can be altered by treatment.

In decision analysis, data on the expected benefits and harm of treatment in patients both with and without the disease (obtained from observational studies or randomized clinical trials), data on the prevalence of disease, and data on the performance of a diagnostic test are combined, thereby allowing comparison of several strategies. In *chapters 7* and *15* decision analysis was used to assess the diagnostic work-up of tubal pathology and work-up of patients with suspected ectopic pregnancy, respectively. An alternative for

clinical decision analysis is the performance of clinical studies, preferably randomized clinical trials, in which the relevant outcome measures after treatment are related to the diagnostic tests under study.

18.4.1 Assessment of the clinical value of diagnostic tests with clinical decision analysis

The general concepts of clinical decision analysis have been introduced in the early 1980s.[25 26] Traditionally, clinical decision analysis contains six steps. First, a decision tree is constructed containing realistic combinations of diagnostic and therapeutic tools. Subsequently, probabilities are assigned for uncertain events. Ideally, estimates of these probabilities are obtained from clinical studies on diagnosis and therapy. In the third phase, utilities are assigned to possible outcomes of the decision tree. Pauker and Kassirer, who developed a model that could help in the decision to perform angiography in the diagnosis of pulmonary embolism, expressed the outcome of each strategy in terms of expected life years. In the evaluation of tubal subfertility (*chapter 7*) the utility of interest was the expected occurrence of pregnancy, whereas in the evaluation of strategies to diagnose ectopic pregnancy (*chapter 15*) the relevant outcome was (prevention of) tubal rupture.

In the fourth step, the probabilities assigned in step two and the utilities assigned in step three are combined by 'folding back' the decision tree. This way, an expected value can be calculated for each strategy. In the fifth step we can then select the decision that has the highest expected utility. The final step is to perform a sensitivity analysis in which the assigned probabilities and utilities are varied within clinically reasonable boundaries to assess the robustness of the outcome of the decision analysis.

Although decision analysis as described above seems a simple and powerful tool for the evaluation of diagnostic test, there are several limitations to its use. First, data on the probabilities of diagnostic and therapeutic processes many uncertain events are often either lacking or conflicting. Second, the relevant outcomes of the evaluated strategies are often multi-dimensional. For a patient with suspected ectopic pregnancy, for example, the side effects of possible treatments, time to diagnosis, number of false-positive diagnosis, tubal rupture, future fertility and costs are all relevant outcomes. If one strategy was superior on all possible outcomes, the choice would be straightforward. Unfortunately, in many decision analyses there is not a single strategy that is superior, as each strategy is often found to have its own advantages and disadvantages. The choice between the various possible outcomes should therefore be made in an early stage of the decision analysis, preferably when constructing the decision tree or assigning the utilities. In some clinical problems, individual preferences might influence the choice for one or the other strategy. In the management of ectopic pregnancy, for example, tubal preservation might be important for patients with future child wish, whereas patients without child wish are only interested in curation of their ectopic pregnancy.

A third problem of clinical decision analysis is that in clinical practice the distinction between presence and absence of a disease is not always clear. For example, does a patient who is subfertile since four years have a disease? Does it matter if there were any abnormalities found at the diagnostic work-up, such as tubal pathology? And does the

fact that treatment-independent pregnancy occurred despite the fact that laparoscopy had shown one-sided tubal occlusion indicate that she is cured from the disease, i.e., that the disease is not present anymore? For ectopic pregnancy the same appears to be the case, although the difference might be more subtle. Does a patient with an ectopically nidated pregnancy and a serum hCG level of 1,400 IU/L has the disease "ectopic pregnancy"? And is this disease also present in case the serum hCG concentration is 200 IU/L or 5,400 IU/L?

These examples illustrate that the border between being healthy and having a disease is often unclear. For the evaluation of diagnostic tests this means that it is often difficult, if not impossible, to define which patients should be classified as diseased and which patients should be classified as healthy.

In this context, one should also realize that making a distinction between healthy and diseased is not the main purpose of diagnostic tests. The main purpose of diagnostic tests is to distinguish persons who are expected to benefit from treatment from persons who are expected not to benefit from treatment. This implicates that it is sometimes of no importance that patients with a mild form of the disease, that does not require treatment, are not detected by a diagnostic test. Since from a pathophysiological point of view these patients really have the disease, it is more appropriate to use the term 'target condition' in stead of the term 'disease' in the evaluation of diagnostic tests. A 'target condition' is in this context a disease status in which the expected benefits of treatment outweigh the expected harm. In contrast, patients not expected to benefit from treatment are considered not to have a target condition. Consequently, diagnostic tests should focus on the detection of persons with the target condition, and not on the detection of disease in a pathophysiologic way.

The knowledge that patients with the target condition will benefit from treatment should be obtained in other clinical studies, preferably randomized clinical trials or prognostic studies. However, this knowledge could also be based on pathophysiologic reasoning. Evaluation of a particular diagnostic test should therefore always start with an inventory of potential treatments. In presence of a treatment of which the effectiveness has been proven in randomized clinical trials, the target condition is then defined by the in- and exclusion criteria that were used in the randomized clinical trial. For a treatment of which the effectiveness has not been established in randomized clinical trials, the target condition is defined by the set of indications that is used for that particular treatment.

An example is the treatment of ectopic pregnancy. Before the introduction of methotrexate as a possible treatment for ectopic pregnancy, the diagnosis of ectopic pregnancy was made at laparoscopy or laparotomy, and the patient could be cured with salpingectomy or salpingostomy. Consequently, ectopic pregnancy confirmed at laparoscopy or laparotomy was used as target condition in diagnostic studies for ectopic pregnancy. In those studies, patients with low serum hCG concentrations were often managed expectantly, which is reasonable since the probability of ectopic pregnancy in these patients was rather low, and since it will always become clear which of these patients has an ectopic pregnancy that requires treatment.

Introduction of systemic methotrexate in the treatment of ectopic pregnancy could lead to changes in the definition of the target condition, which might have consequences for the diagnostic work-up. In *chapter 13* the capacity of non-invasive diagnostic tools to distinguish patients with a ruptured tubal pregnancy from patients with an unruptured tubal pregnancy was evaluated. Before the introduction of systemic methotrexate in the treatment of ectopic pregnancy, this distinction was without clinical relevance, since all patients were treated surgically, independent of the presence of tubal rupture. Consequently, presence of tubal rupture was always detected at explorative surgery, and further surgical management could be adjusted according to the findings.

However, in a non-invasive treatment strategy with systemic methotrexate it is of imminent importance to rule out the presence of tubal rupture and active bleeding before the start of treatment. In terms of benefits and harm, one can say that in case all patients undergo surgery, each patient with ectopic pregnancy, be it with or without tubal rupture, is expected to benefit more from surgery than from no surgery. Since methotrexate can only be applied safely in case of unruptured ectopic pregnancy, patients with unruptured ectopic pregnancy are expected to benefit more from methotrexate than from surgery, whereas patients with ruptured ectopic pregnancy are more likely to benefit from surgery than from methotrexate. Thus, the treatment of unruptured ectopic pregnancy with systemic methotrexate has introduced a new diagnostic category that is of importance for the diagnostic work-up of suspected ectopic pregnancy.

If the performance of a diagnostic test in the detection of a target disorder is known, and if the effectiveness of a treatment for that particular disorder is known, the expected value of the test in a population with a known prevalence can be determined. As was shown in § 18.2.1, one should be aware of the fact that the performance of a test can vary if differences in prevalences are accompanied by difference in characteristics of patients with and patients without the target condition.

18.4.2 Assessment of the clinical value of diagnostic tests in randomized clinical trials

In absence of sufficient evidence on the effectiveness of treatment, studies evaluating diagnostic tests should incorporate evaluation of therapy in their design. In recent years, several diagnostic tests have been evaluated in a randomized setting.[27-29] In such trials, patients were randomly allocated either to undergo the test under study, or not to undergo the test. The value of the test was subsequently determined by comparing the outcome measures of interest in such patients.

Figure 6A shows the design of such trials. Patients randomized not to undergo the test are supposed not to be treated. On the other hand, patients randomized to undergo the test under study are supposed to be treated if the result of the test under study is positive, and supposed not to be treated if the result of the test under study is negative. The difference between the two groups depends on the outcome of three processes. First, the test under study must identify those patients who are expected to benefit from treatment. Second, the clinicians' decision to offer treatment must be in accordance with the result of the test, i.e., a positive result of the test must always be followed by treatment and a negative test result must never be followed by treatment, and the patient must comply

with the offered treatment. Third, the treatment that is performed in patients with a positive test result must be effective.

That the latter may not always be taken into account in the design and interpretation of randomized clinical trials evaluating diagnostic tests becomes clear from the recent controversy on the post-coital test (PCT), that even reached the headlines of national newspapers in The Netherlands. A trial had been performed, in which subfertile couples were randomized either to undergo the PCT or not.[27] No difference was found in pregnancy rates between the couples that had a PCT done and couples that had not, despite the fact that fewer patients were offered treatment in the 'intervention' group as compared to the control group. The authors concluded that the PCT was of no value in the work-up for subfertility. However, clinicians participating in this multi-center trial did not receive a protocol on how to manage a couple with a positive or a negative PCT, and how to manage a couple that did not underwent a PCT. Thus, although a difference was detected in the amount of treatment performed in two groups, it was not clear if the allocation of treatment was in accordance with the test-result.

In a meta-analysis of 12 randomized clinical trials on the effectiveness of doppler-ultrasonography in high-risk pregnancies, Alfirevic and Neilson reported a reduction of the perinatal death rate by 38% (95% confidence interval 15% to 55%).[28] The authors noticed an important lack of explicit decision pathways in the 10 out of the 12 incorporated trials. In fact, the biggest reduction in perinatal death was seen in one of the two trials with a strict management protocol.[30] It can be hypothesized that the difference in effectiveness between two diagnostic strategies increases in case the strategies are applied according to strictly formulated rules, whereas the difference decreases in absence of such rules.

Van Loon *et al.* recently tried to evaluate the value of magnetic resonance pelvimetry in breech presentation.[29] In their randomized clinical trial, 235 women with breech presentation at term were subject to magnetic resonance pelvimetry. They were randomly assigned to two groups. In one group the result of the result of magnetic resonance pelvimetry was disclosed to the treating obstetrician, whereas in the other group the obstetricians remained unaware of the result of this test. Manual pelvimetry, being an alternative test for magnetic resonance pelvimetry, was allowed in the control group, and the management in that group was based on the obstetrician's judgement. In contrast to the trial on the PCT, in this trial there was a strict protocol on the decision for elective caesarian section or trial of labor. The authors reported comparable sectio caesarian rates in both groups, but the number of emergency sectio caesarians was higher in the control group.

However, despite the strict protocol on interpretation of the magnetic resonance pelvimetry, the clinician was aware of the patient's allocation during labor. Thus, it is not unlikely that the obstetrician was more inclined to vaginal delivery when reassured by the magnetic resonance pelvimetry.[31]

Whereas the authors used a design as shown in Figure 6A, it might have been more appropriate to use a design as shown in figure 6B. Consider a woman that is found to have a normal pelvis at magnetic resonance imaging. In case a woman would have been

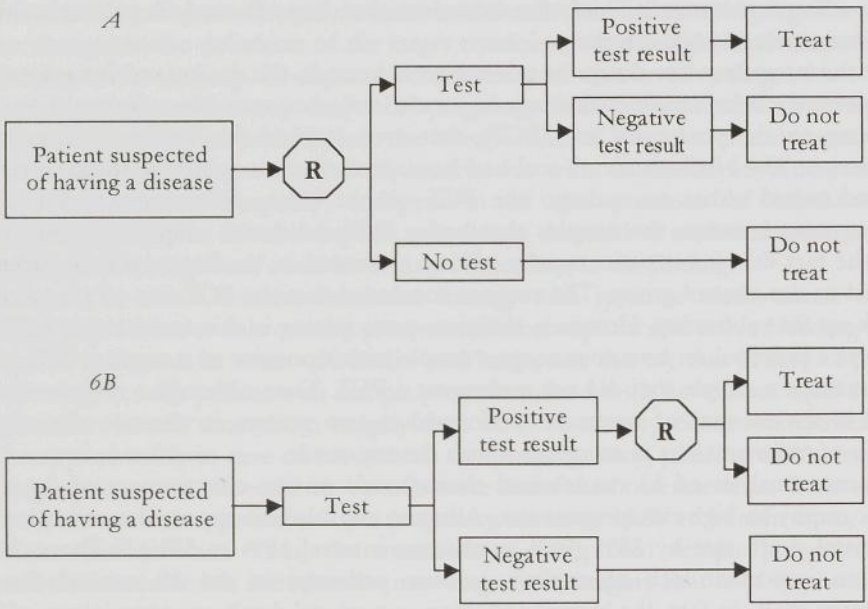

Figure 6AB: Design of a randomized clinical trial, comparing a diagnostic test with no testing. Figure 6A shows the situation in which all patients are randomized to undergo the test. Figure 6B shows a perhaps more efficient design in which all patients undergo the test, and subsequently, only in case of test-positivity, are randomized for treatment or no treatment.

randomly allocated to the magnetic resonance pelvimetry group, the obstetrician would have been reassured by the finding of a normal pelvis at magnetic resonance imaging, and a decision for caesarian section would have been delayed. In case the same woman would have been allocated to the other group, the obstetrician would have been inclined to perform caesarian section at an earlier stage of labor, thereby increasing the rate of emergency caesarian sections. Indeed, the authors reported that the duration of the first stage of labor in the magnetic resonance pelvimetry group that underwent eventually an emergency sectio caesarian, was almost 3 hours longer than in the control group.[32]

In contrast, a woman with an abnormal pelvis at magnetic resonance pelvimetry would have undergone primary caesarian section when she would have been randomized to undergo magnetic resonance pelvimetry, whereas she would have been set up for vaginal delivery in case she would have been allocated to the other group, conditionally on a normal pelvis at manual examination. It is the latter category of women that is thought to benefit from magnetic resonance pelvimetry. Thus, a more appropriate design of the trial would have been to randomize only those women, in whom magnetic resonance pelvimetry and manual pelvimetry had shown discordant test results. Such a design is

shown in Figure 6B. In such a design the first category of women (normal pelvis both at manual and magnetic resonance pelvimetry) would not have been randomized at all, thereby preventing that a different inclination for caesarian section depending on disclosure of the result of magnetic resonance pelvimetry would have interfered with the association of interest.

When considering a randomized clinical trial for the assessment of diagnostic tests, it is important to realize that the difference between the two groups in a randomized clinical trial evaluating diagnostic tests is completely explained by the group of patients that would have had discordant test results, would they have undergone both tests. In a randomized clinical trial in which a patient would have undergone only one of the two tests under study, we would never have known in which of the patients the test results would have been discordant. It is likely that an alternative design, in which both tests are performed in each patient, but only patients with discordant test results are randomized, is more efficient (Figure 6B).[33] Such a trial would, in case costs of trial monitoring and treatment are relatively high as compared to costs of the tests under study, reduce the total costs. Moreover, it would reveal insight in the type of patients in which the two tests would disagree. Such a design could not only be used in case two tests are compared, but also in case performance of a test is compared with no testing at all. In that case, it would be sufficient to randomize the patients with a positive or a negative test, depending on the fact if one would withhold or perform treatment in the group of patients not undergoing the test. In the evaluation of the PCT, for example, it would have sufficed to perform the PCT in all patients, and then randomize only those patients who had an abnormal PCT either to undergo treatment or not to undergo treatment.

18.5 Towards a framework for the evaluation of diagnostic tests.

The previous paragraphs have shown the importance of taking into account relevant patient outcomes when assessing the clinical value of a diagnostic test. Several authors have incorporated outcomes relevant for the patient in frameworks that can be used in the evaluation of diagnostic tests.[34] Yet, clinical guidelines for the use of diagnostic tests without considering patient outcomes are provided by many authors.[35] The purpose of this paragraph is to provide a framework that can be used for the evaluation of the effectiveness of diagnostic tests.

Figure 7 shows such a framework. The first phase deals with the development of a diagnostic test. Fundamental research in the field of physics, biochemistry, biology, physiology or psychology will lead to the development of new tests that have potential for clinical practice. After the development phase, the test has to be reproducible and has to show some discriminative performance during 'in vitro' studies. These two phases are beyond the scope of this thesis.

After a diagnostic test has been developed and after it has shown to be reproducible and to have some discriminative performance in the laboratory, it should be evaluated in a clinical context. When addressing the clinical application of the test, it is important to determine if the purpose of the test is either to reduce uncertainty in order to support a

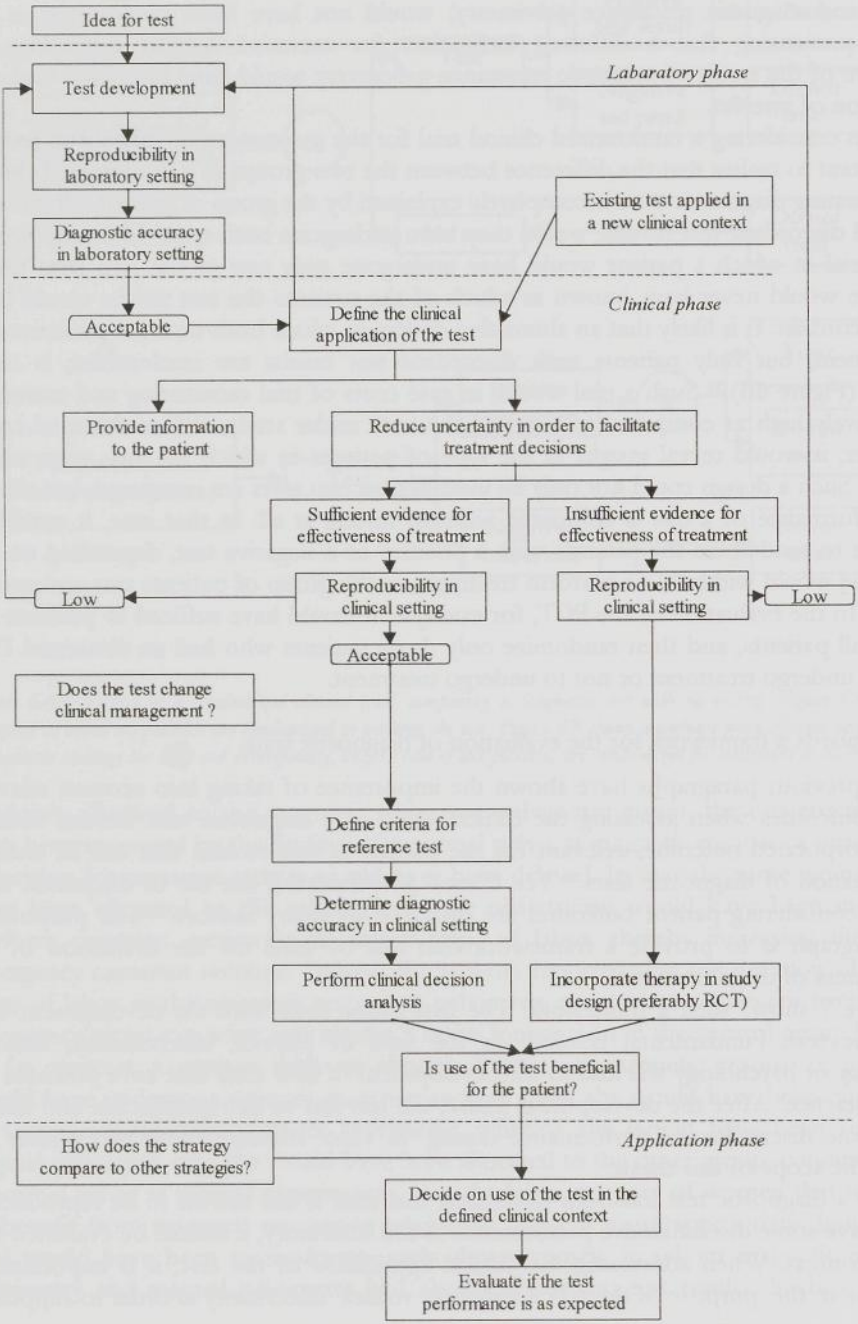
Idea for test
Labaratory phase
Test development
Reproducibility in laboratory setting
Diagnostic accuracy in laboratory setting
Existing test applied in a new clinical context
Clinical phase
Acceptable
Define the clinical application of the test
Provide information to the patient
Reduce uncertainty in order to facilitate treatment decisions
Sufficient evidence for effectiveness of treatment
Insufficient evidence for effectiveness of treatment
Low
Reproducibility in clinical setting
Reproducibility in clinical setting
Low
Acceptable
Does the test change clinical management ?
Define criteria for reference test
Determine diagnostic accuracy in clinical setting
Perform clinical decision analysis
Incorporate therapy in study design (preferably RCT)
Is use of the test beneficial for the patient?
Application phase
How does the strategy compare to other strategies?
Decide on use of the test in the defined clinical context
Evaluate if the test performance is as expected

Figure 7 (opposite page): Framework for the evaluation of a diagnostic test

decision on treatment, or if its only purpose is to provide information to the patient. As stated previously, assessment of the value of such information is beyond the scope of this thesis.

If the purpose of the test is to reduce uncertainty on the disease status in order to support decisions on treatment, information on the effectiveness of therapy of the target condition should be collected, preferably before the study that is evaluating the diagnostic test is designed. One has to be informed about both the effectiveness of treatment and the characteristics of the patients in whom the treatment was found to be effective. If there appears to be an effective treatment (preferably evaluated in randomized clinical trials), the diagnostic study should focus on the detection of patients with baseline characteristics corresponding with those used in such trials, i.e., in which treatment is known to be effective. The 'target condition' that is used as reference test in studies evaluating a diagnostic test should correspond with the disease in which the treatment was found to be effective. In case data on the effectiveness of therapy are lacking, studies evaluating a diagnostic test have to incorporate therapy. For such studies, randomized designs as discussed in § 18.3.2 would suffice. In such studies it is of imminent importance that interpretation of test results and subsequent therapeutic measures are clearly defined.

Before a study assessing the diagnostic accuracy of a diagnostic test or a study assessing a diagnostic test in a randomized setting are started, it would be useful to be informed about the reproducibility of the test in the relevant clinical context. As was shown in § 18.2.4, reproducibility is one of the key determinants of the diagnostic performance of a test. If the reproducibility of a test is found to be low, the performance of a diagnostic test can never be good. In *chapter 2*, we started the evaluation of HSG with a study of the reproducibility of its interpretation. The fact that the reproducibility of the diagnosis of adhesions was found to be moderate already indicated that this diagnosis can not be made very accurately with HSG. In chapter 13, the meta-analysis summarizing 13 studies that reported on the diagnosis of peritubal adhesions with HSG as compared to laparoscopy confirmed the limited potential of HSG in the diagnosis of adhesions.

Once the accuracy of a test has been established in a clinical setting or once the effectiveness of a test has been assessed in a randomized clinical trial, the question arises if that test should be used in clinical practice, and in which patients it might be useful. The first issue that is of importance when assessing this question is whether performance of the test changes clinical management. If performance of a test would lead to a change of clinical policy in a very limited number of patients, the probability that the test will be useful is likely to be low. The second issue that should be addressed is the performance of the test as compared to other tests. When addressing this question it is essential that the sequence at which tests are performed in clinical practice are taken into account. For example, data of medical history are easily available after the performance of a medical history. The question that arises is how many other tests contribute to the information that is already available.

18.6 Summary and conclusions

The main aim of this chapter was to discuss how to evaluate the effectiveness of diagnostic tests. In § *18.2* the traditional framework for the evaluation of diagnostic tests was discussed. The well-known parameters sensitivity and specificity were explained, as well as LRs and ROC-curves. All these parameters express the performance of diagnostic tests. In theory, these parameters are independent of the prevalence of disease. Consequently, they can be combined with the prevalence of disease in a particular setting, in order to calculate positive and negative predictive values, which determine the interpretation in clinical practice.

In § *18.2.1* it was subsequently demonstrated that differences in prevalence of disease, if they are associated with patient characteristics, might influence the sensitivity and/or specificity of a diagnostic test. This means that sensitivity and specificity are not always constant in populations with different prevalences of disease. Thus, the impact of patient characteristics on sensitivity and specificity should be incorporated in the evaluation of diagnostic tests.

In § *18.2.2* it was demonstrated that the accuracy of diagnostic tests can be determined in case-control studies or in cohort studies. Cohort studies assemble patients at risk for a disease, in whom both the test under study and the reference standard are performed, whereas case-control studies assemble cases with the disease and controls without the disease. Case-control studies are therefore likely to overestimate the performance of the test under study as compared to cohort studies.

In § *18.2.3* several forms of bias were discussed that can occur in studies that evaluate the performance of a test. In § *18.2.4* the relation between the reproducibility of a particular test and its performance was explained. When the reproducibility of a test is poor, the accuracy of the test can not be good, thereby making the test useless for clinical practice.

§ *18.2.5* dealt with meta-analysis of diagnostic tests. Heterogeneity in studies reporting on the performance of a diagnostic test might be due to differences in cutoff levels that are used in different studies. If this were to be the case, higher sensitivity would be accompanied by a lower sensitivity and vice versa. Subsequently, the performance of the test can be expressed by estimating a summary ROC-curve.

§ *18.3* described the differences between screening and diagnosis. Screening was defined as systematic detection of disease before the onset of symptoms by means of a relatively simple test. The essential difference between diagnosis and screening is that in case of diagnosis the patient contacts the health care provider, usually after the onset of complaints, whereas in case of screening the health care provider contacts the person who is at risk for having a disease or a preliminary phase of it.

§ *18.4* dealt with the assessment of the clinical value of diagnostic tests. The primary purpose of a diagnostic test is to identify patients that should be treated, since treatment is the primary way to improve the health status of these patients. Thus, the value of a diagnostic test can therefore only be assessed by taking into account subsequent consequences of treatment, in terms of health outcomes that are relevant for the patient.

Is the therapeutic context of the test under study defined and are these appropriately treatments evaluated?	
♦ *Yes, the therapeutic context of the test under study is defined and treatments are appropriately evaluated?*	♦ *No, the therapeutic context of the test under study is not defined?*
• Were indications for treatment incorporated in the definition of the reference standard?	• Were treatment alternatives incorporated in the design of the study?
• Were data sufficient for calculation of sensitivity, specificity and LRs?	• Was the study randomized?
• Was, in case of a test with a continuous test-result, the performance of the test assessed with a ROC-curve?	• Were the outcome measures relevant for clinical practice?
• Was the impact of patient characteristics on the discriminative capacity of the test evaluated?	• Were results of the test under study interpreted according to a strict management protocol?
• Were the data collected in a case-control study or a cohort study?	
• Was the data-collection affected by bias?	
✓ Was the decision to perform the reference test (partially) dependent on the result of the test under study (verification bias)?	
✓ Was inclusion of a patient in the study dependent on the result of the test under study (selection bias)?	
✓ Was the test under study used as part of the reference test (incorporation bias)?	
✓ Was the person interpreting the reference test aware of the result of the test under study (diagnostic review bias)?	
✓ Was the person interpreting the test under study aware of the result of the reference test (test review bias)?	
• Were patients with inconclusive test result or patients in whom the test failed incorporated in the analysis?	
• What is the value of the test if its results are incorporated in a decision analysis?	

Table 3: Checklist for the critical appraisal of articles reporting on a diagnostic test.

Apart from the accuracy of a test as expressed by sensitivity, specificity or LRs, the value of a test also depends on the prevalence of disease and on the effectiveness and side- effects of the available treatment options(s). If the prevalence of a disease is either very high or very low, it is unlikely that performance of a test will be useful. Similarly, in absence of an effective treatment for a certain disease, it is of limited value to detect such a disease. Finally, costs and side-effects of the test itself are of importance.

In § *18.4.1* it was shown that the clinical value of diagnostic tests can be evaluated with clinical decision analysis. A model can be built in which the accuracy of a diagnostic test in the distinction between those who should be treated and those who should not be treated is combined with the effectiveness of treatment. Subsequently, the expected value of multiple diagnostic strategies can be evaluated for different prevalences of disease. Such models can also take into account costs and side-effects

In § *18.4.2* it was demonstrated that randomized clinical trials are an alternative for the use of decision analysis in the evaluation of diagnostic tests. Randomized clinical trials can be useful in the evaluation of diagnostic tests, especially in absence of information on the effectiveness of relevant treatments. However, the use of randomized trials in the evaluation of diagnostic tests has several pitfalls. First, the clinicians' decision to offer treatment must be in accordance with the result of the test under study, i.e., a positive result of the test must always be followed by treatment and a negative test result must never be followed by treatment, and the patient must comply with the offered treatment. Second, the treatment that is performed in patients with a positive test result must be effective. Several examples of randomized trials that assessed diagnostic tests were discussed.

Finally, § *18.5* provided a framework that can be used for the evaluation of diagnostic tests. Key issue in this framework is the availability of sufficient evidence on the effectiveness of relevant treatment. In the presence of such evidence, it is sufficient to evaluate how accurate the test under study can identify the target condition, i.e., the disease status that can be improved by treatment. The value of the test can subsequently be determined by modeling. In absence of such evidence, the evaluation of treatment can be incorporated in the study, preferably a randomized clinical trial.

Table 3 provides a checklist that can be used in the critical appraisal of studies that report on the effectiveness of diagnostic tests. The first question to be answered is whether the therapeutic context of the test under study is defined and whether the effectiveness of such treatments is already properly evaluated. If this is the case, one should first verify if the target condition as defined in the diagnostic study corresponds with the in- and exclusion criteria as used in the studies on therapy.

Subsequently, data sufficient for calculation of sensitivity, specificity and LRs must be available from the articles, if they are not provided in the paper. In case of a test with a continuous test result, the performance of the test should be expressed with ROC-curves. The validity of the study must be assessed by defining the type of study and checking for potential biases, as described in § *18.2.2* and § *18.2.3*. Finally, the value of the test must be determined in a decision analysis, that is preferably supplied in the diagnostic study, although in absence of such an analysis it might be necessary that some 'quick and dirty' analysis is performed by the clinician-reader.

In absence of evidence on the effectiveness of therapy, one should evaluate whether treatment alternatives are incorporated in the design of the study. If treatment alternatives were taken into account, the next question is if the relevant outcome measures were used.

As shown in *§ 18.4.2*, the results of the test under study must be interpreted according to a strict management protocol.

18.7 References

1. Feinstein AR. The architecture of clinical research. WB Saunders company. Philadelphia, 1985.
2. Kraemer HC. Evaluating Medical Tests. SAGE Publications Newbury Park, London, New Delhi, 1992.
3. Goldschlager N, Selzer A, Cohn K. Treadmill stress tests as indicators of presence and severity of coronary artery disease. Ann Int Med 1976;85:277-86.
4. Rozanski A, Diamond GA, Berman D, Forrester JS, Morris D, Swan HJ. The declining specificity of exercise radionuclide ventriculography. N Engl J Med 1983;309:518-22.
5. Osbakken MD, Okada RD, Boucher CA, Strauss HW, Pohost GM. Comparison of exercise perfusion and ventricular function imaging: an analysis of factors affecting the diagnostic accuracy of each technique. J Am Coll Card 1984;3:272-83.
6. Moons KGM, Vanes GA, Deckers JW, Habbema JDF, Grobbee DE. Limitations of sensitivity, specificity, likelihood ratios, and Bayes theorem in assessing diagnostic probabilities - a clinical example. Epidemiol 1997;8:12-7.
7. Lijmer JG, Mol BWJ, Heisterkamp S, Bonsel GJ, Prins MH, Van der Meulen J, Bossuyt PMM. Empirical evidence of design-related bias in diagnostic studies. Submitted
8. Mol BWJ, Bairam N, Lijmer JG, Wiegerinck MAHM, Bongers MY, Van der Veen F, Bossuyt PMM. The accuracy of CA-125 in the diagnosis of endometriosis: a meta-analysis. Fertil Steril 1998;70:1101-8.
9. Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. Biometrics 1983;39:207-15.
10. Quinn MF. Relation of observer agreement to accuracy according to a two-receiver signal detection model of diagnosis. Med Decis Making 1989;9:196-206.
11. Fleiss JL. Statistical methods for rates and proportions. 1981. New York. John Wiley and Sons.
12. Chalmers TC, Lau J. Changes in clinical trials mandated by the advent of meta-analysis. Stat Med 1996;15:1263-8.
13. Ioannidis JP, Cappelleri JC, Lau J. Issues in comparisons between meta-analyses and large trials. JAMA 1998;279:1089-93.
14. Irwig L, Tosteson A, Gatsonis C, Lau J, Colditz G, Chalmers TC, Mosteller F. Guidelines for meta-analyses evaluating diagnostic tests. Ann Inter Med 1993;120:667-76.
15. Kardaun JPWF, Kardaun OJWF. Comparative diagnostic performance of three radiological procedures for the detection of lumbar disk herniation. (1990) Thesis. Erasmus University. Rotterdam. The Netherlands.
16. Moses LE, Shapiro D, Littenberg B. Combining independent studies of a diagnostic test into a summary ROC-curve: Data-analytic approaches and some additional considerations. Stat Med 1993;12:1293-1316.
17. Ter Riet G, Kleijnen J, Knipschild P. Acupuncture and chronic pain: a criteria based meta-analysis. J Clin Epidemiol 1990;43:1191-9.
18. Mol BWJ, Lijmer JL, Ankum WM, Van der Veen F, Bossuyt PMM. Accuracy of single serum P measurement in the diagnosis of ectopic pregnancy; a meta-analysis. Hum Reprod 1998;13:3220-7.
19. Hakama M. Screening. In: Holland WW. Detels R, Knox G. Oxford textbook of public health, 2nd ed. Oxford: Oxford University Press, 1991.
20. Mackenbach JP. Screening: nieuwe mogelijkheden, nieuwe controversen? Ned Tijdschr Geneesk 1995;139:734-9.
21. Jaeschke R, Guyatt G, Sackett DL. Users' Guides to the Medical Literature III. How to use an article about a diagnostic test. A. Are the results of the study valid? JAMA 1994;271:703-7.

22. Berwick DM, Weinstein MC. What do patients value? Willingness to pay for ultrasound in normal pregnancy. Med Care 1985;23:881-93.
23. Asch DA, Patton JP, Hershey JC. Knowing for the sake of knowing: The value of prognostic information. Med Decis Making 1990;10:47-57.
24. Marteau TM, Kidd J, Cuddeford L, Walker P. Reducing anxiety in women referred for colposcopy using an information booklet. Br J Health Psych 1996;1:181-9.
25. Pauker SG, Kassirer JP. The threshold approach to clinical decision making. N Engl J Med 1980;302:1109-17.
26. Weinstein MC, Fineberg HV, Elstein AS, Frazier HS, Neuhauser D, Neutra RR, McNeil BJ. Clinical Decision Analysis. WB Saunders Company. 1980. Philadelphia London Toronto.
27. Oei SG, Helmerhorst FM, Bloemenkamp KW, Hollants FA, Meerpoel DE, Keirse MJ. Effectiveness of the postcoital test: randomized controlled trial. BMJ 1998;317:502-5.
28. Alfirevic Z, Neilson JP. Doppler ultrasonography in high-risk pregnancies: systematic review with meta-analysis. Am J Obstet Gynecol 1995;172:1397-87.
29. Van Loon AJ, Mantingh A, Serlier EK, Kroon G, Mooyaart EL, Huisjes HJ. Randomized controlled trial of magnetic resonance pelvimetry in breech presentation at term. Lancet 1997;350:1799-804.
30. Almstrom H, Axelsson O, Cnattingius S, Ekman G, Maesel A, Ulmsten U, Arstrom K, Marsal K. Comparison of umbilical artery velocimetry and cardiotocography for surveillance of small-for-gestational-age-fetuses. Lancet 1992;340:936-40.
31. Van der Post JAM, Maathuis JB. Magnetic resonance pelvimetry in breech presentation. Letter-to-the-editor. Lancet 1998;351:913.
32. Van Loon AJ, Mantingh A. Magnetic resonance pelvimetry in breech presentation. Reply. Lancet 1998;351:913.
33. Bossuyt PMM, Tijssen JGP. Randomization in the evaluation of diagnostic procedures. Presented at the 10th annual meeting of the international society for technology and health care. 1994. Baltimore USA.
34. Guyatt GH, Tugwell PX, Feeny DH, Haynes RB, Drummond M. A framework for clinical evaluation of diagnostic technologies. CMAJ 1986;134:587-94.
35. Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research. Getting better but still not good. JAMA 1995;274:645-51.