Part 2

Principles of Clinical Research
INTRODUCTION

A 55-year-old man visits his general practitioner (GP) complaining of dyspepsia. He has had these complaints for more than 3 months, but their frequency and severity have increased over the last 4 weeks. The patient has a history of angina but has not required sublingual nitroglycerin for more than 2 years. He is known to the GP as having been unsuccessful in quitting smoking despite frequent attempts to do so. The GP asks several additional questions related to the nature and severity of the dyspepsia to estimate the chance that the patient suffers from a peptic ulcer. The GP also asks about possible anginal complaints. A short physical examination reveals nothing except some epigastric discomfort during palpation of the abdomen. The GP considers a peptic ulcer the most likely diagnosis. The probability of a coronary origin of the complaints is deemed very low. The GP decides to test for Helicobacter pylori serology, to further increase (rule in) or decrease (rule out) the probability of (H. pylori-related) peptic ulcer. The H. pylori test is negative. The GP prescribes an acid-suppressing agent and asks the patient to visit again in a week. When the man visits the GP again, his complaints have virtually disappeared.

DIAGNOSIS IN CLINICAL PRACTICE

Doctors devote much of their time to diagnosing diseases in patients presenting with particular symptoms or signs. Determining a diagnosis for a patient is important because it provides insight into the prognosis of the patient and directs the physician in making decisions for appropriate patient management (see Box 2–1).

The diagnostic process in daily practice typically starts with a patient presenting a certain complaint—symptom or sign—that makes the practitioner suspicious of him or her having a particular disorder (target disease).
out of a series of possible disorders (differential diagnosis) [Sackett et al., 1985]. The target disease can best be viewed as the disorder at which the diagnostic process is initially targeted, either because it is the most serious of the possible diagnoses (“the one not to miss”) or, a priori, the most probable one. During the diagnostic process, the physician first estimates the probability, or likelihood, of the presence of the target disease in view of the alternative diagnoses (including the absence of any disease) based on information obtained through history taking, including knowledge about a patient’s individual and family medical history, and physical examination. This diagnostic probability estimation is typically an implicit and subjective process (see Box 2–2).

In addition to clinical data about the patient, nonclinical data such as age, gender, and working conditions also may be considered. The estimated probability of the target disease will guide the doctor in choosing the most appropriate action. The physician may perform additional diagnostic tests, initiate therapeutic interventions, or, perhaps most importantly, may decide to refrain from further diagnostic or therapeutic actions for that disease (e.g., when the probability of that disease is considered low enough) and possibly search for other underlying diseases [Ferrante di Ruffano et al., 2012]. The diagnostic workup is a continuing process of updating the probability of the target disease presence given all available documented information on the patient. The goal of this workup is to achieve a relatively high or a relatively low probability of a certain diagnosis, that is, the threshold probability beyond or below which a doctor is confident enough about the presence or absence of a certain diagnosis to guide clinical decisions. Threshold probabilities are determined by the consequences of a false-positive or false-negative diagnosis. These critically depend on the anticipated course or prognosis of the diagnosis considered and the potential beneficial and adverse effects of possible additional diagnostic procedures or treatments. Importantly, these two thresholds, A and B, are commonly implicitly defined in daily practice and will often vary between doctors. Often, history taking and physical examination already provide important and sufficient diagnostic

---

BOX 2–1  Quotation About Clinical Judgment

Knowing how to live with uncertainty is a central feature of clinical judgment: the skilled physician has learned when to take risks to increase certainty and when to simply tolerate uncertainty.

—Riegelman, 1990

information to rule in or rule out a disease with enough confidence so that the estimated probability of presence of the disease is below A or above B (see Figure 2–1).

But when the probability of the disease is estimated to lie in the grey middle area (between A and B), additional diagnostic tests are commonly ordered to decrease the remaining uncertainty about the presence or absence of the disease. Typically, this additional testing first includes simple, easily available tests such as blood and urine markers or simple imaging techniques like chest x-ray. If after these tests are conducted doubt remains (i.e., the probability of disease presence has not yet crossed the thresholds A or B), more invasive and costly diagnostic procedures are applied such as magnetic resonance imaging (MRI), computed tomography (CT), or positron emission tomography (PET) scanning, arthroscopy, and/or biopsy. This process of diagnostic testing ends when the estimated probability of the target disease becomes sufficiently higher or lower than the A or B threshold to guide medical action.
In the example of our patient with complaints of dyspepsia, history taking and physical examination apparently did not provide the doctor with enough information to decide about the initiation of therapeutic interventions, for example, symptomatic treatment with acid-suppressing agents or, alternatively, triple therapy to treat an underlying *H. pylori* infection. In view of the patient burden of invasive *H. pylori* testing (i.e., gastroscopy with biopsy) in combination with the relatively mild complaints and potential benefits of *H. pylori*–targeted therapy, the physician decided to perform a noninvasive serology test, although this test is considered less accurate than the gastroscopy. Apparently, the negative test results indeed convinced the physician that the probability of *H. pylori* ulcer disease was lower than the clinically relevant threshold (e.g., 10 or 20%) because triple therapy targeted at *H. pylori* was not initiated. Instead, symptomatic treatment, an acid-suppressing drug, was prescribed. The probability of coronary heart disease—one of the differential diagnoses of a patient with these complaints—as the underlying cause of the complaints also was consid-

![Diagram: Figure 2-1 Diagnostic Testing](image-url)
erved to be very low from the start (far below threshold A for this disease), such that no additional tests for that diagnosis were ordered.

This example may seem subjective, not quantitative and not evidence based, but the diagnostic process in clinical practice is often just like that. In contrast to many therapeutic interventions, quantitative evidence of the value of diagnostic tests and certainly of the added value of a test beyond previous, more simple test results, is often lacking [Linnet et al., 2012; Moons et al., 2012c]. Given the importance of diagnosis in everyday practice, there is an urgent need for research providing such quantitative knowledge [Grobbee, 2004; Knottnerus, 2002b].

The diagnostic process thus is a multivariable concern. It typically involves the documentation and interpretation of multiple test results (or diagnostic determinants), including nonclinical patient information [Moons et al., 1999]. In practice, hardly any diagnoses are based on a single diagnostic test. The number of diagnostic tests applied in everyday practice may differ considerably and depends, for example, on the targeted disease, patient characteristics, and the diagnostic certainty required to decide on patient management (see Box 2–3). Importantly, a natural hierarchy of testing exists. Almost without exception, any diagnostic workup starts with BOX 2–3

Primum Non Nocere

*Primum non nocere* (first do no harm) refers to the principle that doctors should always take into account the possible harm of their actions to patients, and that an intervention with an obvious potential for harm should not be initiated, notably when the benefits of the intervention are small or uncertain.

Although this Hippocratic principle is most often applied to discussions on the effects of therapeutic interventions, it is equally applicable to diagnostic tests, especially for the more invasive and burdensome ones. When the course of management for a patient has been determined, additional diagnostic tests obviously have no benefit and can therefore only be harmful, albeit sometimes to the healthcare budget only. In daily practice many diagnostic tests are being performed that have no potential helpful consequences for patient management. Especially when additional test ordering is relatively easy, for example, serum parameters and imaging such as x-rays, the potential consequences for patient management, as well as possible harm, are not always taken into account. In a patient with a rib contusion as a result of a fall, an x-ray to rule out a rib fracture is useless, because the test result will not influence treatment (i.e., rest and painkillers). The challenge to the physician in any diagnostic process thus not only lies in choosing the optimal diagnostic tests and in what order, but also in knowing when to stop testing.
nonintrusive tests such as history taking and physical examination. Although one could argue about whether these should be considered “tests,” we will treat them as such here, as each consecutive finding will influence the probability of disease, just as a blood test would. This is followed by simple laboratory or imaging tests, and eventually more burdensome and expensive tests, such as imaging techniques requiring contrast fluids or biopsies. Subsequent test results are always interpreted in the context of previous diagnostic information [Moons et al., 1999; Moons & Grobbee, 2002a]. For example, the test result “presence of chest pain” is obviously interpreted differently in a healthy 5-year-old girl than in a 60-year-old man with a history of myocardial infarction. The challenge to the physician lies in predicting the probability of the absence or presence of a certain target disease based on all documented test results. This requires knowledge about the contribution of each test result to the probability estimation. The diagnostic value of the H. pylori test in the earlier example is negligible if it adds nothing to the findings offered by the few minutes of history taking and physical examination, information that is always acquired by physicians anyway. More technically, the H. pylori test is worthless if the test result does not change (increase or decrease) the probability of presence of peptic ulcer disease as based on the results from history taking and physical examination. Importantly, in case the next step in clinical management is already decided upon (when the disease probability is below A or above B in Figure 2–1), one may, and perhaps should, refrain from additional testing.

The works of the 18th century Scottish pastor and mathematician, Thomas Bayes, have been instrumental in the development of a more scientific approach toward the diagnostic process in clinical practice. He established a mathematical basis for diagnostic inference. Bayes recognized the sequential and probabilistic nature of the diagnostic process. He emphasized the importance of prior probabilities, that is, the probability of the presence of a target diagnosis before any tests are performed. He also recognized that, based on subsequent test results, doctors will update this prior probability to a posterior probability. The well-known Bayes’ rule formally quantifies this posterior probability of disease presence given the test result, based on the prior probability of that disease and the so-called diagnostic accuracy estimates (such as sensitivity and specificity or likelihood ratio) of that test (see Box 2–4). Although it has repeatedly been shown that this mathematical rule often does not hold—because the underlying assumption of constant sensitivity and specificity or likelihood ratio across patient subgroups is not realistic in most situations [Detrano et al., 1988; Diamond, 1992; Hlatky et al., 1984; Moons et al., 1997]—the rule has been crucial in understanding
the probabilistic and stepwise nature of diagnostic reasoning in clinical practice.

We should emphasize that setting a diagnosis is not itself a therapeutic intervention. It is a vehicle to inform patients and guide patient management [Biesheuvel et al., 2006; Bossuyt et al., 2012]. An established diagnosis is a label that, despite being highly valued by medical professionals, is of no direct consequence to a patient other than to obtain a first estimate of the expected course of the complaints and to set the optimal management strategy. Accordingly, a diagnostic test commonly has no direct therapeutic effects and therefore does not directly influence a patient’s prognosis. Once a diagnosis, or rather the probability of the most likely diagnosis, is established and an assessment of the probable course of disease in the light of different treatment alternatives (including no treatment) has been made, the optimal treatment strategy will be chosen to eventually improve the patient’s prognosis. There are also other pathways through which a diagnostic test may affect a patient’s health [Ferrante di Ruffano et al., 2012]. Knowledge of specific test results or disease presence may change the patient’s (and the physician’s) expectations and perceptions, and test results may shorten the time between symptom onset and treatment initiation, as well as improve treatment adherence. Finally, a diagnostic test may have direct therapeutic properties and change patient outcomes. Such procedures are rare, but salpingography to determine patency of the uterine tubes is an example.

Finally, the difference between diagnosing and screening for a disease should be recognized. The former starts with a patient presenting with a particular symptom and sign suspected of a particular disease and is inherently multivariable. Screening for a disease typically starts with asymptomatic individuals and is commonly univariable. Examples include phenylketonuria screening in newborns and breast cancer screening in middle-aged women, where a single diagnostic test is performed in all subjects irrespective of symptoms or signs. In this chapter, we will deal with diagnosing exclusively.

FROM DIAGNOSIS IN CLINICAL PRACTICE TO DIAGNOSTIC RESEARCH

Diagnostic research should be aimed at improving the diagnostic process in clinical practice. Typically it focuses on identifying combination(s) of tests that have the largest diagnostic yield. In clinical epidemiologic terms, the occurrence relation of diagnostic research predicts the probability of the
Example of a Two-by-Two Table with Test Results and Bayes’ Rule

Test characteristics of test (T) N-terminal pro B-type natriuretic peptide (NT-proBNP; cut-off 36 pmol/L) in the detection of heart failure in primary care patients with conditions known to be associated with a high prevalence of heart failure.

<table>
<thead>
<tr>
<th></th>
<th>NT-proBNP positive (T+)</th>
<th>NT-proBNP negative (T–)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heart failure present (D+)</td>
<td>9</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>Heart failure absent (D–)</td>
<td>69</td>
<td>55</td>
<td>124</td>
</tr>
<tr>
<td>Total</td>
<td>78</td>
<td>55</td>
<td>133</td>
</tr>
</tbody>
</table>

where positive predictive value = P(D+ | T+) = 9/78 = 12%; negative predictive value = P(D– | T–) = 55/55 = 100%; sensitivity = P(T+ | D+) = 9/9 = 100%; specificity = P(T– | D–) = 55/124 = 44%; likelihood ratio positive test (LR+) = P(T+ | D+)/[1 – P(T– | D–)] = (9/9)/(69/124) = 1.8; likelihood ratio negative test (LR–) = [(1 – P(T+ | D+))/P(T– | D–)] = (0/9)/(55/124) = 0.

Bayes’ rule:

\[
P(D+ | T+) = \frac{P(D+) \cdot P(T+ | D+)}{P(D+) \cdot P(T+ | D+) + P(D–) \cdot P(T+ | D–)} \quad \text{(Eq. 1)}
\]

\[
P(D+ | T+) = \frac{P(D+) \cdot \text{sensitivity}}{P(D+) \cdot \text{sensitivity} + [1 - P(D+)] \cdot (1 - \text{specificity})} = \frac{9/133 \cdot 1}{9/133 \cdot 1 + 124/133 \cdot 0.56} = 0.12 = 12%
\]

and

\[
P(D– | T+) = \frac{P(D–) \cdot P(T+ | D–)}{P(D+) \cdot P(T+ | D+) + P(D–) \cdot P(T+ | D–)} = \quad \text{(Eq. 2)}
\]

\[
P(D– | T+) = \frac{P(D–) \cdot (1 - \text{specificity})}{P(D+) \cdot \text{sensitivity} + [1 - P(D+)] \cdot (1 - \text{specificity})} = \frac{124/133 \cdot 0.56}{9/133 \cdot 1 + 124/133 \cdot 0.56} = 0.88 = 88%
\]
Alternative (so-called odds) notation of Bayes’ rule → (1) divided by (2):

\[
\frac{P(D^+ | T^+)}{P(D^- | T^+)} = \frac{P(D^+)}{P(D^-)} \cdot \frac{P(T^+ | D^+)}{P(T^+ | D^-)} = \frac{P(D^+)}{P(D^-)} \cdot LR^+ \rightarrow 
\]

Posterior odds \( (D^+ | T^+) \) = prior odds \( (D^+ | T^-) \) * LR^+

Note: Odds\( (D^+) = \frac{P(D^+)}{1 - P(D^+)} \]

For sequential diagnostic tests, Bayes’ rule theoretically can be simply extended:

\[
\frac{P(D^+ | T_1^+, T_2^+, T_3^+)}{P(D^- | T_1^+, T_2^+, T_3^+)} = LR(T_1^+) \cdot LR(T_2^+) \cdot LR(T_3^+)
\]

Note that this form of Bayes’ rule assumes that the results of test 1 to test 3 are independent of each other. However, it has been shown that this assumption in practice typically does not hold, as test results are often mutually related simply because they are reflections of the same underlying disease (see text).
presence of the disease of interest as a function of multiple diagnostic determinants in the relevant domain. The domain is defined by patients suspected of having that particular disease. Diagnostic determinants are the diagnostic tests under study (so-called index tests) and typically include findings from history taking (including age, gender, symptoms, and known comorbidity) and physical examination (signs), and if applicable and necessary, the findings from more advanced diagnostic testing.

The diagnostic process and thus diagnostic research is predictive or descriptive by nature, as its object is prediction of the presence of the yet unknown underlying disease. The goal is not to explain the cause of the disease under study. Consequently, confounding variables (i.e., factors that may distort a causal relationship between a particular causal determinant and an outcome) do not play a role in diagnostic research and are not part of the occurrence relation. This is in sharp contrast to causal research, where confounders are of critical importance. In diagnostic research, all other determinants merely serve as additional diagnostic test results that may be helpful in further distinguishing between those with and without the disease. Importantly, diagnostic research should be performed in close adherence to daily clinical practice to ensure the applicability of the findings. Thus, the typical features of the diagnostic process outlined previously should be taken into account in the design of the study. This has important consequences for the choice of, for example, the study population, the diagnostic determinants to be evaluated, their hierarchy and temporal sequence of measurement, and the data analysis.

**DIAGNOSTIC RESEARCH VERSUS TEST RESEARCH**

Alas, many published diagnostic studies are better characterized as test research than as diagnostic research. The objective of test research is to assess whether a single diagnostic test (index test) adequately can show the presence or absence of a particular disease [Linnet et al., 2012]. Often these studies include a group of patients with the target disease and a group of patients without this disease in whom the results of the index test are also measured. Typically, the results of the index test are categorized as positive or negative and the study results are summarized in a \(2 \times 2\) contingency table (Box 2–4). The table allows for calculation of the four classic measures to estimate diagnostic accuracy in test research. These are:

1. Positive predictive value \([P(D+ | T+)]\); probability (P) of the presence of disease in those with a positive test result
2. Negative predictive value \([P(D-|T-)]\); probability of absence of disease in those with a negative test result.
3. Sensitivity \([P(T+|D+)]\); probability of a positive test given presence of the disease (the true positive rate)
4. Specificity \([P(T-|D-)]\); probability of a negative test in those without disease (the true negative rate)

Other—though less often applied—parameters include the likelihood ratio of a positive test (i.e., the probability of a positive test in the diseased divided by the probability of a positive test in the nondiseased), the likelihood ratio of a negative test (i.e., the probability of a negative test in the diseased divided by the probability of a negative test in the nondiseased), or the odds ratio (which can be calculated as the ratio of the former two). The latter is seldom applied, but it is occasionally used in diagnostic meta-analyses [Reitsma et al., 2012]. If the index test results are not dichotomous but measured on a continuous scale, receiver operating characteristic (ROC) curves can be produced, based on sensitivity and specificity of the different cut-off values of the diagnostic test to be evaluated [Hanley and McNeil, 1982; Harrell et al., 1982].

Test research as described here often deviates from the main principle of clinically relevant diagnostic research in that clinical practice is not followed, first and foremost because the diagnostic process by definition involves multiple tests and a natural hierarchy of diagnostic testing. Second, test research often does not include representatives of the relevant patient domain, that is, patients presenting with symptoms and signs suggestive of the target disease. Rather, a group of patients with evident disease is selected and compared to a group of nondiseased patients, sometimes even healthy individuals who are obviously not suspected of the disease under study. Such selection of study subjects, however, will lead to biased estimates of the test’s performance.

There is a clear difference in the occurrence relation between test research and diagnostic research. The occurrence relation of test research can be described as:

\[
P(T) = f(D)
\]

where \(P(T)\), that is, the probability (0–100%) of the test result of the single index test \(T\), is studied as a function of the presence or absence of the target disease \(D\).

In the case of a dichotomous test, this occurrence relation can be rewritten for the estimation of sensitivity as:

\[
P(T+) = f(D+)
\]
and for estimation of the specificity as:

\[ P(T-) = f(D-) \]

where \( T+ \) and \( T- \) indicate a positive and negative index test result, respectively, and \( D+ \) and \( D- \) the presence or absence of the target disease.

The occurrence relation of test research that focuses on predictive values of a single test can be summarized as:

\[ P(D) = f(T) \]

where the probability of the presence of disease (\( P[D] \)); range (0–100%) is studied as a function of the test result.

In case of a dichotomous test, this occurrence relation can be rewritten for the estimation of the positive predictive value as:

\[ P(D+) = f(T+) \]

and for estimation of the negative predictive value as:

\[ P(D-) = f(T-) \]

The occurrence relation of diagnostic research (i.e., clinically relevant diagnostic studies including multiple diagnostic tests) can be summarized as:

\[ P(D) = f(T_1, T_2, T_3, \ldots T_n) \]

where \( T_1 \) to \( T_n \) are the consecutive multiple diagnostic index tests (or determinants) being studied.

A study to determine the value of plasma N-terminal pro B-type natriuretic peptide (NT-proBNP) levels in diagnosing heart failure serves as an example of a diagnostic study primarily presented as test research. NT-proBNP, a neuropeptide produced in the human cardiac ventricle as a result of increasing pressure, was assessed in a sample of 133 primary care patients [Hobbs et al., 2002]. Selection of these patients was based on the presence of a condition known to be associated with a higher prevalence of heart failure (i.e., a history of myocardial infarction, angina, hypertension, or diabetes), and the study was not restricted to the clinically more relevant group of patients presenting with complaints (e.g., fatigue or dyspnea) suggestive of heart failure.

Box 2–4 summarizes the main results. In addition, Bayes’ rule, calculating the post-test probability (or odds) of disease as the product of the pre-test probability (or odds) of the disease, and the test’s likelihood ratio are illustrated using the data derived from this study.
It was concluded from the study that NT-proBNP has value in the diagnosis of heart failure. Its main use would be to rule out heart failure in patients with suspected heart failure in whom normal concentrations of NT-proBNP are found. Several critical remarks can be made about this study, most of which were recognized by the authors. First, the focus of this study on the NT-proBNP test as a single test to diagnose or rule out heart failure does not reflect the diagnostic approach in clinical practice. NT-proBNP will never be applied as the sole diagnostic test in diagnosing heart failure. Simpler diagnostic tools inevitably are used first, notably information on age, sex, comorbidity, and symptoms and signs, before additional tests, such as NT-proBNP and perhaps electrocardiography, or even echocardiography, are applied [Rutten et al., 2005b]. The clinically more relevant research aim would thus be to assess whether NT-proBNP appreciably adds to the diagnostic information (such as signs and symptoms) that is readily available in clinical care.

This can only be achieved by comparing the diagnostic performance of two diagnostic strategies: one including all diagnostic information available to the physician before NT-proBNP measurement is executed, and one including the same information plus the NT-proBNP levels. In doing so, the multivariable nature of the diagnostic process in clinical practice is taken into account as well as the inherent hierarchy of diagnostic testing. The authors indeed performed a multivariable logistic regression analysis to determine whether a model including sex, history of myocardial infarction or diabetes, Q waves, or bundle branch block pattern on the electrocardiogram, and NT-proBNP performed better in diagnosing heart failure than a similar model excluding NT-proBNP. Nonetheless, the added diagnostic value of NT-proBNP was not emphasized in the presentation of the results or in the conclusion, nor were symptoms and signs included as possible diagnostic tests in the multivariable analysis. In addition, the study population can be criticized. The ability of a diagnostic test or combination of tests to distinguish between diseased and nondiseased should be studied in those patients in whom the diagnostic problem truly exists. In other words, the patients should be representatives of the domain of patients suspected of having that disease and in whom the physician will consider diagnostic testing. This is crucial because the value of diagnostic tests critically depends on the patient mix (see Box 2–5) [Lijmer et al., 1999; Rutjes et al., 2006]. Most patients included in the NT-proBNP study (i.e., mainly patients with conditions known to be associated with a high prevalence of heart failure) were not representative of the clinically relevant domain of patients visiting their primary care physician with symptoms and signs suggestive of
Sensitivity and Specificity Are Not Constant

Are sensitivity and specificity given properties of a diagnostic test, and do predictive values critically depend on the prevalence of the disease?

The common emphasis on sensitivity and specificity in the presentation of diagnostic studies is at least partly attributable to the notion that predictive values critically depend on the population studied, whereas sensitivity and specificity are considered by many to be constant [Moons & Harrell, 2003]. There is no doubt that predictive values of diagnostic tests are influenced by the patient domain. This may be best illustrated by comparing the performance of a test in primary and secondary care. Because of the inherent higher prevalence of the relevant disease in suspected patients in secondary care compared to primary care (because of the selection of patients with a higher probability of disease for referral), positive predictive values are commonly higher in secondary care (i.e., fewer false-positives) than in primary care (more false-positives), while negative predictive values are usually higher in primary care (fewer false-negatives). Sensitivity, specificity, and likelihood ratios indeed are not directly influenced by the prevalence of the disease because these parameters are conditional upon the presence or absence of disease. It has been shown extensively, however, that they do vary according to differences in the severity of disease [Hlatky et al., 1984; Detrano et al., 1988; Diamond, 1992]. In secondary care, for example, where more severely ill patients will be presented than in primary care, higher levels of diagnostic markers of a particular disease (and thus more test positives) can be expected among those with the disease than in primary care. This will result in a higher sensitivity in secondary care than in primary care and a higher specificity in primary care [Knottnerus, 2002a]. That sensitivity and specificity are not constant is illustrated in two studies by the same researchers on the value of near patient testing for Helicobacter Pylori infection in dyspepsia patients. The sensitivity and specificity in the primary care setting were 67% and 98%, respectively, while these values were 92% and 90% in secondary care [Duggan et al., 1999; Duggan et al., 1996].

heart failure. Thus, the applicability of the findings to patients encountered in daily practice is limited.

The focus on the quantification of the value of a single test to diagnose or rule out a disease and the common preoccupation of such research with a test’s sensitivity and specificity are typical of prevailing diagnostic research [Moons et al., 2004a; Moons et al., 2012c]. This is also illustrated by
the following statements found in classic textbooks in clinical epidemiology or biostatistics:

Identify the sensitivity and specificity of the sign, symptom, or diagnostic test you plan to use. Many are already published and sub specialists worth their salt ought either to know them from their field or be able to track them down [Sackett et al., 1985].

and

For every laboratory test or diagnostic procedure there is a set of fundamental questions that should be asked. Firstly, if the disease is present, what is the probability that the test result will be positive? This leads to the notion of the sensitivity of the test. Secondly, if the disease is absent, what is the probability that the test result will be negative? This question refers to the specificity of the test [Campbell & Machin, 1990].

As the goal of determining a diagnosis for patients is to estimate the probability of disease given the diagnostic test results, the parameters of interest undoubtedly are the posterior probabilities or predictive values, which directly reflect the diagnostic probabilities needed for decision making in clinical practice. Indeed, patients do not enter a physician’s office saying, “I have been diagnosed with this particular disease and would like to know the probability that the available tests are positive.” For the doctor, this probability (i.e., sensitivity) is similarly uninformative. A focus on probabilities of test results given the presence or absence of disease—sensitivity and specificity—is unjustified from a clinical point of view. It should be emphasized that in the NT-proBNP study example, the authors stated that their main conclusion (that heart failure can be excluded in those with normal NT-proBNP values) was indeed based on the excellent negative predictive value. In our experience, researchers as well as journal editors are reluctant to dismiss the sensitivity and specificity as the most important parameters in diagnostic research, as is also reflected in guidelines on the reporting [Bossuyt et al., 2003a, 2003b] and critical appraisal of diagnostic studies [Whiting et al., 2011], although more recently methods focusing on predictive values have been advocated [Leeflang et al., 2012; Reitsma et al., 2012].

A first step to de-emphasize these measures when judging the value of diagnostic tests is to change the order in which the traditional parameters are presented and to present predictive values first [Moons & Harrell, 2003; Rutten et al., 2006]. Diagnostic knowledge is not provided by answering the question, “How good is this test?” Diagnostic knowledge is the information needed to answer the question, “What is the probability of the presence or absence of a specific disease given these test results?”
Notwithstanding its limitations, test research—focusing on estimating the accuracy of a single test—may offer relevant information. Most notably, it is helpful in the developmental phase of a new diagnostic test, when the accuracy of the test is yet unknown. One will often first assess whether the test provides different results in those with overt disease and those without disease, sometimes even using healthy control subjects [Fryback & Thornbury, 1991; Linnet et al., 2012]. Furthermore, test research can be valuable in the realm of screening for a particular disorder in asymptomatic individuals. In this context, no test results other than the single screening test are considered. Depending on the type of screening not even age and gender may need to be accounted for [Moons et al., 2004a].

**DIAGNOSTIC RESEARCH**

Because the object of diagnosis in practice is to predict the probability of the presence of disease from multiple diagnostic test results, the design of diagnostic research is very much determined by the understanding, if not mimicking, of everyday practice [Moons & Grobbee, 2005]. In the following sections, the three components of clinical epidemiologic diagnostic study design will be discussed: theoretical design, design of data collection, and design of data analysis.

**Theoretical Design**

As mentioned earlier, the occurrence relation of diagnostic research is:

\[ P(D) = f(T_1, T_2, T_3, \ldots T_n) \]

The domain of the occurrence relation in diagnostic research typically includes patients suspected of a particular disease, usually defined by the presence (or combination) of particular symptom(s) and/or sign(s) that have led to consultation of a physician. In this context, the research objective can be to assess the optimal diagnostic strategy; that is, to determine which combination of diagnostic determinants in what order most adequately estimate the probability of disease presence. The goal can also be to assess whether a certain, often newly developed, diagnostic test provides additional diagnostic value in clinical practice. *Added value* means in addition to currently available or previously applied diagnostic tests. This implies a comparison of two occurrence relations: one excluding and one including the new test in the list of determinants studied. Furthermore, one could aim to compare two tests or different combinations of tests, for example,
when a new less burdensome or more inexpensive test serves as an alternative to another established diagnostic test. This implies a comparison of an occurrence relation with the routinely available test(s) (including this established test) and a second occurrence relation including the same tests, except that this established test is substituted by the alternative or new test.

For most diagnostic studies, showing that a particular diagnostic test, combination of tests, or test strategy improves estimation (prediction) of the presence of a disease is enough from a clinical point of view, because the clinical consequences (i.e., targeted therapy) and the effects of such therapy are well established [Bossuyt et al., 2012]. Showing that NT-proBNP clearly improves the ability to diagnose or exclude heart failure in suspected patients may suffice to apply such a test in daily practice as there is an impressive body of evidence showing that targeted treatment in heart failure improves survival and quality of life [Kelder et al., 2011]. Sometimes, however, the therapeutic consequences of a diagnosis may not be clear, such as when a new test provides truly novel disease information that potentially calls for other treatment choices compared to the currently available test. An example is functional imaging with PET in diagnosing pancreatic cancer, for which CT is the widely accepted reference standard. Compared to CT, PET may be especially helpful in detecting smaller lesions and distant metastases. Application of PET may lead to other diagnostic classifications that would require initiating other treatment options that potentially have different patient outcomes than the use of CT [Lord et al., 2006; Moons, 2010]. In such a situation, studies may be conducted to estimate the additional benefit of a new diagnostic test or strategy on the patient's prognosis (e.g., in terms of morbidity, mortality, or quality of life), rather than doing a study comparing the diagnostic accuracy of PET with CT as the reference standard. Although inspired by a diagnostic question, such studies are not simply predictive. They become analogous to studies assessing the effects of therapeutic interventions on patient outcome and, consequently, carry the characteristics of intervention research. In intervention research the aim is to explain (“Does addition of this test cause an improvement in patients’ prognoses?”) rather than to predict (“Does this test improve the estimation of the probability of the presence of a certain disease?”). Thus, confounding becomes an issue, because one wishes to ensure that the observed effects are indeed attributable to the diagnostic test or strategy. All of this has important consequences for the theoretical design (notably for the outcome definition), the design of data collection (where randomized trials with a relevant time horizon may be an attractive option), and the data analysis (see Box 2–6). For the purpose of understanding the principles of clinical
Illustration of the difference between (typical) diagnostic research, assessing the contribution of multiple diagnostic determinants to the estimation (prediction) of the presence of a certain disease and diagnostic intervention research aimed at estimating (in this case explaining) the effect of diagnostic tests (plus subsequent interventions) on the patient’s prognosis. The latter type of research becomes intervention research, and requires taking extraneous determinants (i.e., confounders) into account.

**Diagnostic Research**

$$P(D) = f(T_1, T_2, T_3, \ldots T_n)$$

Where $P(D)$ is the probability of the presence (i.e., prevalence) of the disease of interest and $T_1, T_2, T_3, \ldots T_n$ represent the diagnostic determinants to be assessed.

The occurrence relation of diagnostic research covers the bold part of this scheme:

**Diagnostic problem** $\rightarrow$ **diagnostic strategy** $\rightarrow$ **diagnosis** $\rightarrow$ intervention $\rightarrow$ outcome

**Diagnostic Intervention Study**

$$\text{Prognostic outcome} = f[(T_1, T_2, T_3, \ldots T_n) + (I|ED)]$$

Where the prognostic outcome could be any clinically relevant patient outcome, such as survival, incidence of a specific outcome, duration of the complaints or quality of life; $T_1, T_2, T_3, \ldots T_n$ represent the diagnostic determinants to be assessed; I is intervention following diagnosis and ED are extraneous determinants (or confounders) that should be taken into account in this causal study.

The occurrence relation of a diagnostic intervention study covers this entire scheme:

**Diagnostic problem** $\rightarrow$ **diagnostic strategy** $\rightarrow$ **diagnosis** $\rightarrow$ intervention $\rightarrow$ outcome
epidemiologic study design in typical diagnostic research (where index tests are compared to a reference standard in the relevant patient domain) this category of diagnostic intervention studies is not addressed in detail in this chapter; some other texts discuss the principles of diagnostic intervention studies [Biesheuvel et al., 2006; Bossuyt et al., 2000; Bossuyt et al., 2012; Lijmer & Bossuyt, 2009].

Design of Data Collection

Time
The object of the diagnostic process is cross-sectional by definition. In diagnostic research the probability of the presence of a disease (prevalence) is estimated, not its future occurrence. Accordingly, the data for diagnostic studies are collected cross-sectionally. The determinant(s) (the diagnostic test results) and the outcome (the presence or absence of the target disease as determined by the so-called reference standard) are theoretically determined at the same time. This is the moment that the patient presents with the symptoms or signs suggestive of the disease \( t = 0 \). Even when the assessment of all diagnostic determinants to be studied takes some time and when it takes several days or weeks before the definitive diagnosis becomes known, this time period is used to determine whether at \( t = 0 \) the disease was present. Also, when a “wait and see” period of several months (e.g., to see whether an underlying disease, such as cancer, becomes manifest or whether targeted therapy has a beneficial effect) is used to set the final diagnosis, these additional findings are used to establish the diagnosis present at the time the patient presented the symptoms (i.e., at \( t = 0 \)) [Reitsma et al., 2009]. Thus, in our view, diagnostic research is cross-sectional research (time is zero). It should be noted, however, that others consider time to be larger than zero when it takes some time to set the final diagnosis and, as a consequence, they characterize the design of data collection as a follow-up or cohort study.

Census or Sampling
Generally, diagnostic research takes a census approach in which consecutive patients suspected of a certain disease and who fulfill the predefined inclusion criteria are included. The potentially relevant diagnostic determinants as well as the “true” presence or absence of the target disease are measured in all patients.

Sometimes, however, a sampling approach (i.e., a case-control study; see the later chapter on case-control studies) can offer a valid and efficient
alternative. In a diagnostic case-control study (which is a cross-sectional case-control study), all patients suspected of the target disease who are eventually diagnosed with the disease (“cases”) are studied in detail, together with a sample of those suspected of the disease who turn out to be free from the target disease (“controls”). This implies that the outcome (reference standard) has to be assessed in all patients suspected of the target disease (otherwise the cases cannot be identified and the controls cannot be sampled), but that the diagnostic determinants only have to be measured in cases and controls. As in diagnostic research using a census approach, the goal is to obtain absolute probabilities of disease presence given the determinants. Consequently, in the data analysis of a diagnostic case-control study, the sampling fraction of the controls should always be accounted for. A diagnostic case-control study offers a particularly attractive option when the measurement or documentation of one or more of the diagnostic tests under study are time consuming, burdensome to the patient, or expensive, such as certain imaging tests [Rutjes et al., 2005]. Diagnostic case-control studies are still relatively rare, despite their efficiency [Biesheuvel et al., 2008a]. In the example in Box 2–7, a case-control approach was chosen to assess the added value of cardiac magnetic resonance (CMR) imaging in diagnosing heart failure in patients known to have chronic obstructive pulmonary disease. Because of the costs, time, and patient burden involved, CMR measurements were performed in all patients with heart failure (cases) but in only a sample of the remainder of the participants (controls) [Rutten et al., 2008].

Confusingly, diagnostic studies comparing test results in a group of patients with the disease under study—often those in an advanced stage of disease—with test results in a group of patients without this disease, often a group of healthy individuals from the population at large, tend to be referred to as diagnostic case-control studies [Rutjes et al., 2005]. Many of these studies are not case-control studies, however, as there is no sampling of controls from the study base [Biesheuvel et al., 2008a]. In addition, as discussed earlier, such studies will bias the estimates of diagnostic accuracy of the tests being studied and compromise the generalizability of the study results. This is because the cases and certainly the healthy controls do not reflect the relevant patient domain, which is all those suspected of having the disease for whom the tests are intended.

**Experimental or Observational**

Diagnostic research is typically *observational* research. In patients suspected of the disease in daily practice, the diagnostic determinants of interest (most of which will be measured in clinical practice anyway), including
BOX 2–7 Example of a Diagnostic Case-Control Study

**BACKGROUND:** Although cardiovascular magnetic resonance (CMR) imaging is well established, its diagnostic accuracy in identifying chronic heart failure (CHF) in patients with chronic obstructive pulmonary disease (COPD) has not yet been quantified.

**METHODS:** Participants were recruited from a cohort of 405 patients aged 65 years or older with mild to moderate and stable COPD. In this population, 83 (20.5%) patients had a new diagnosis of CHF, all left-sided, established by an expert panel using all available diagnostic information, including echocardiography. In a nested case-control study design, 37 consecutive COPD patients with newly detected CHF (cases) and a random sample of 41 of the remaining COPD patients (controls) received additional CMR measurements. The value of CMR in diagnosing heart failure was quantified using univariable and multivariable logistic modeling in combination with area under the receiver operating characteristic curves (ROC area).

**RESULTS:** The combination of CMR measurements of left-ventricular ejection fraction, indexed left- and right-atrial volume, and left-ventricular end-systolic dimensions provided high added diagnostic value beyond clinical items (ROC area = 0.91) for identifying CHF. Left-sided measurements of CMR and echocardiography correlated well, including ejection fraction. Right-ventricular mass divided by right-ventricular end-diastolic volume was higher in COPD patients with CHF than in those without concomitant CHF.

**CONCLUSIONS:** Easily assessable morphologic and volume-based CMR measurements have excellent capacities to identify previously unknown left-sided chronic heart failure in mild to moderate COPD patients. There seems to be an adaptive tendency to concentric right-ventricular hypertrophy in COPD patients with left-sided CHF.


possible new tests, will be measured and the presence of disease will be determined using the reference standard. Such a cross-sectional study will be able to show which combination of tests best predicts the presence of disease or whether a new test improves diagnostic accuracy.

As discussed, setting a diagnosis is not an aim in itself, but rather a vehicle to guide patient management and treatment in particular. The ultimate
goal of diagnostic testing is to improve patient outcomes. Hence, it has widely been advocated that when establishing the accuracy of a diagnostic test or strategy, its impact on patient outcomes also must be quantified. Consequently, it has been proposed that experimental studies (diagnostic intervention studies comparing two diagnostic strategies) be used to answer diagnostic research questions [Bossuyt et al., 2012; Lord et al., 2006].

If a cross-sectional diagnostic study has indicated that the diagnostic test or strategy improves estimation of the presence of the disease, the effect on patient outcome can usually be validly established without the need for a diagnostic intervention study [Koffijberg et al., 2013]. After all, earlier studies often adequately quantified the effects on patient outcome of the available treatment(s) for that disease. Using simple statistical or decision modeling techniques, one can combine the results of the cross-sectional diagnostic accuracy study and those of randomized therapeutic intervention studies. Hence, the effect on patient outcome can be quantified if (1) diagnostic research has shown that the diagnostic test or strategy improves diagnostic accuracy and (2) the effects of available therapeutic interventions in that disease on patient outcome are known, preferably from randomized trials. An example in which a randomized study was not necessary to quantify the effect of the new test on patient outcome is a study assessing whether an immunoassay test for the detection of H. pylori infection can replace the established but more costly and invasive reference test (a combination of rapid urease test, urea breath test, and histology) [Weijnen et al., 2001]. The new test indeed provided similar diagnostic accuracy. As consensus exists about the therapeutic management of patients infected with H. pylori (based on randomized controlled trials establishing the efficacy of treatment [McColl, 2002]), a subsequent diagnostic intervention study to quantify the effects of using the new immunoassay test on patient outcome was not needed.

There are situations, however, in which diagnostic intervention studies are needed to properly quantify the consequences of a novel diagnostic test or strategy on patient outcome [Biesheuvel et al., 2006; Bossuyt et al., 2000; Lord et al., 2006]. Notably, when a new diagnostic technology under study might be “better,” to the extent that it provides new information potentially leading to other treatment choices, than the existing tests or strategy, a randomized trial may be useful. As described previously, functional imaging with PET in diagnosing pancreatic cancer, for which CT is the current reference, is an example. Also, when there is no direct link between the result of the new diagnostic test under study and an established treatment indication, such as the finding of uncalcified small nodules (less than 5.0 mm) when screening for lung cancer with low-dose spiral CT scanning, an
experimental approach quantifying the effect on patient outcome may be required. When an acceptable reference standard for a disease is lacking, for instance, in a diagnostic study in suspected migraine or benign prostatic hyperplasia, a diagnostic intervention may also be the best option. Finally, as mentioned earlier, the index test itself (e.g., salpingography in suspected tubal blockage) may have direct therapeutic effects.

When performing a diagnostic intervention study to determine the impact of a diagnostic test or strategy on patient outcome, an initial diagnostic research question is transformed into a therapeutic research question (with the goal of establishing causality) with corresponding consequences for the design of the study. A disadvantage of a randomized approach to directly quantifying the contribution of a diagnostic test and treatment to the patient’s outcome is that it often addresses diagnosis and treatment as a single combined strategy, a “package deal.” This makes it impossible to determine afterward whether a positive effect on patient outcome can be attributed solely to the improved diagnostic accuracy or to the new subsequent treatment strategies.

**Study Population**

A diagnostic test or strategy should be able to distinguish between those with the target disease and those without, among subjects representing the relevant clinical domain. The domain is thus defined by patients suspected of having a particular disease. Consequently, patients in whom the presence of disease has already been established or in whom the probability of the disease is considered high enough to initiate adequate therapeutic actions fall outside the domain, similar to when the probability of disease is deemed sufficiently low to exclude the diagnosis (see also Figure 2–1). Furthermore, we recommend that investigators restrict domain definitions, and thus the study population, to the setting or level of care (e.g., primary or secondary care), as the diagnostic accuracy and combinations of these tests usually vary across care settings [Knottnerus, 2002a; Oudega et al., 2005a]. This is a consequence of differences in the distribution of severity of the disease across the different settings.

The population of a study could be defined as all consecutive patients suspected of the disease of interest that present themselves to one of the participating centers during a defined period and in whom the additional diagnostic tests under investigation are considered. Exclusion criteria should be few to ensure wide applicability of the findings. They would typically include alarm symptoms requiring immediate action or referral (e.g., melena in the dyspepsia example in the beginning of this chapter) and contraindications for one of the major diagnostic determinants (tests).
involved (e.g., claustrophobia when MRI assessments are involved). One could argue that “patients suspected of the disease” as an inclusion criterion is too subjective. In many studies the definition, therefore, includes symptoms and signs often accompanying the disease. For example, a study to address the added value of a novel test to diagnose or exclude myocardial infarction in the primary care setting could include “patients with symptoms suggestive of myocardial infarction in primary care.” Alternatively, the study population can be defined as “patients with chest pain or discomfort in primary care” or a combination of the two: “patients with chest pain of discomfort or other symptoms and signs compatible with a myocardial infarction in primary care” [Bruins Slot et al., 2013].

**Diagnostic Determinants**

As the diagnosis in practice is typically made on the basis of multiple diagnostic determinants, all test results that are (potentially) used in practice should be considered and measured. In the earlier example of the *H. pylori* test to diagnose peptic ulcer, the main signs and symptoms as well as the *H. pylori* test have to be included as potential determinants. There is, however, a limit to the number of tests that can be included in a study, because of logistics and the larger sample size required with each additional test that is considered (see the following discussion). Hence, the choice of the determinants to be included should be based on both the available literature and a thorough understanding of clinical practice.

To optimize the applicability of the findings of diagnostic research, the assessment of the diagnostic determinants should resemble the quality of this information in daily clinical practice. Consequently, one could argue that all determinant information should be collected according to usual care, without efforts to standardize or improve the diagnostic assessment. In a study involving multiple sites and physicians, this may significantly increase inter-observer variability in diagnostic testing, which means the potential diagnostic value of test results could be underestimated, although the study would indicate the current average diagnostic value of the tests in clinical practice. This effect is likely to be larger for more subjective tests, such as auscultation of the lungs. An alternative would be to train the physicians to apply a standardized diagnostic assessment. One may also ask experts in the field to do the diagnostic tests under study. This, however, has the disadvantage that it will likely overestimate the diagnostic accuracy of the tests in daily practice and reduce the applicability of the study results. For a multicenter, multi-doctor study, we recommend a pragmatic approach where all diagnostic determinants are assessed as much as possible according
to daily practice and by the practicing physicians involved, with some efforts to standardize measurements.

Outcome

The outcome in diagnostic research is typically dichotomous: the presence or absence of the disease of interest (e.g., myocardial infarction or pneumonia). As discussed, in clinical practice commonly more than one disease is considered in a patient presenting with particular symptoms and signs, that is, the so-called differential diagnosis [Sackett et al., 1985]. Thus, the outcome should be polytomous rather than dichotomous, although in daily practice sequential dichotomous steps are often taken; the most likely (or most severe) disease in the differential diagnosis is diagnosed or excluded before the next diagnosis is considered. Diagnostic research with polytomous or even ordinal outcomes is relatively rare and the data analysis is more complicated [Harrell, 2001]. Current methodologic developments in this field no doubt will increase the use of polytomous outcomes in diagnostic research [Biesheuvel et al., 2008b; Roukema et al., 2008; Van Calster et al., 2012].

In diagnostic research, as in each epidemiologic study, adequate assessment of the outcome is crucial. The outcome should be measured as accurately as possible and with the best available methods. The term most often applied to indicate the ideal diagnostic outcome is gold standard, referring to the virtually nonexistent situation where measuring the disease is devoid of false-negatives and false-positives [Reitsma et al., 2009]. More recently, the more appropriate term reference standard was introduced to indicate the “non-golden” properties of almost all diagnostic procedures in today’s practice, including procedures like biopsy combined with histologic confirmation for cancer diagnoses. Very few diagnostic procedures do not require human interpretation. Deciding on the reference standard is a crucial but difficult task in diagnostic research. The reference standard is the best procedure(s) that exists at the time of study initiation to determine the presence or absence of the target disease. The word best in this context means the measurement of disease that best guides subsequent medical action. Hence, the reference method to be used in a diagnostic study may very well include one or a combination of expensive and complicated tests that are not routinely available or applied in everyday clinical practice. Note that this contrasts with the assessment of the diagnostic determinants of interest, which should more or less mimic daily practice to enhance generalizability of study results to daily practice.

Preferably, the final diagnosis should be established independent of the results of the diagnostic tests under study. Commonly, the observer
who assesses the final diagnosis using the reference method is blinded for all of the test results under study. If this blinding is not guaranteed, the information provided by the preceding tests may implicitly or explicitly be used in the assessment of the final diagnosis. Consequently, the two information sources cannot be distinguished and the estimates of accuracy of the tests being studied may be biased. Although theoretically this bias can lead to both an under- and overestimation of the accuracy of the evaluated tests, it commonly results in an overestimation; the final diagnosis may be guided to some extent by the results of the test under evaluation, artificially decreasing the number of false-positive and false-negative results. This kind of bias is often referred to as diagnostic review or incorporation bias [Begg & Metz, 1990; Ransohoff & Feinstein, 1978; Sackett et al., 1985; Swets, 1988].

The possibility of blinding the outcome assessors for the results of the tests under study depends on the type of reference standard applied. It is surely feasible if the reference standard consists of a completely separate test, for example, imaging techniques or serum levels of a marker. Because this kind of reference test is not available for many diseases (e.g., psychiatric disorders), or is infeasible or even unethical to apply in all cases (notably when the test is invasive and patient burdening), next best solutions are often sought. In particular, an approach involving a so-called consensus diagnosis determined by an outcome panel often is applied; this often is combined with a clinical follow-up period to further promote an adequate assessment of the presence of the disease [Begg, 1990; Reitsma et al., 2009; Swets, 1988]. Outcome panels consist of a usually unequal number of experts on the clinical problem. During consensus meetings, the panel establishes the final diagnosis in each study patient based on as much patient information as possible. This includes information from patient history, physical examination, and all additional tests. Often, any clinically relevant information (e.g., future diagnoses, response to treatment targeted at the outcome disease) from each patient during a prespecified follow-up period is also forwarded to the outcome panel in order to allow for a better judgment on whether the target disease was present at the time of (initial) presentation [Moons & Grobbee, 2002b]. When using a consensus diagnosis based on all available information as the reference standard, the test results studied as potential diagnostic determinants are usually also included (“incorporated”) in the outcome assessment, leading to a risk of incorporation bias. To fully prevent incorporation bias, the outcome panel should decide on the final diagnosis without knowledge of the results of the particular test(s) under study. This may seem an attractive solution, but limiting the information forwarded to the panel may increase misclassification
in the outcome assessment. There are no set solutions to this dilemma that is inherent to using a consensus diagnosis as the reference standard. The pros and cons of excluding or including the results from all or some of the tests under study in the assessment of the final diagnosis by the outcome panel should be weighed in each particular study. Consider a study that aims to assess the diagnostic value of NT-proBNP serum levels or echocardiography in addition to signs and symptoms in patients suspected of heart failure. As in several earlier studies on suspected heart failure, an outcome panel can determine the “true” presence or absence of heart failure [Moons & Grobbee, 2002b; Rutten et al., 2005b]. When studying the accuracy of a test known to receive much weight in the consensus judgment (in this example echocardiography and to a lesser extent NT-proBNP levels), it is preferable not to use these tests in the assessment of the final diagnosis. Doing so requires that the remaining diagnostic information, including clinical follow-up data, enable the panel to accurately diagnose patients. Lack of availability of the NT-proBNP levels will probably not pose a major problem, but withholding the echocardiographic findings, a key element in the diagnosis of heart failure, from the outcome panel may seriously endanger the validity of the outcome assessment. Consequently, we may be able to quantify the added value of NT-proBNP levels but not the added value of the echocardiogram [Kelder et al., 2011]. Alternatively, the outcome panel could judge the presence or absence of heart failure first without considering the echocardiographic findings and then subsequently with the echocardiography results. Comparing the outcome classification according to both approaches may provide some insight into the effect of incorporation bias on the (boundaries of the) accuracy of the test under study, in this case echocardiography.

As mentioned earlier, in certain situations it is not feasible and may even be unethical to apply the best available reference method in all study patients at the time of presentation, in particular when the reference test is invasive and may lead to complications (such as pulmonary angiography in suspected pulmonary embolism). Also in studies in suspected malignancies, it is often difficult to establish or rule out a malignancy at \( t = 0 \), even when multiple tests, including sophisticated imaging techniques, are performed. Under such circumstances, a clinical follow-up period may offer useful information. It should be emphasized here that a clinical follow-up period is applied to assess whether the disease of interest was indeed present at the time of presentation of the complaints (\( t = 0 \)). It is then assumed that the natural history of the (untreated) target disease implies that the target disease was present but unrecognized at \( t = 0 \). A clinical follow-up period to establish a diagnosis has been successfully applied in studies on the accuracy
of diagnostic tests for a variety of diseases, including pulmonary embolism, bacterial meningitis, and certain types of cancer. For example, Fijten et al. [1995] studied which signs and symptoms were helpful in ruling out colorectal cancer in patients presenting with fecal blood loss in primary care. It was impossible to perform colonoscopies and additional imaging or surgery in all participants to rule in or out a malignancy at t = 0. Therefore, all patients were followed for an additional period of at least 12 months after inclusion in the study, assuming that colorectal cancer detected during the follow-up period would indicate presence of the cancer at baseline. Obviously, the follow-up period should be limited in length, especially in diseases with a relatively high incidence, to prevent new cases from being counted as prevalent ones. The acceptable clinical follow-up period varies and depends on the natural history and incidence of the disease studied. A 6- to 12-month period is often encountered in the literature for cancer studies. For venous thromboembolism this is usually 3 months, and in a study of bacterial meningitis it was 1 week.

Besides documenting the natural history of a disease during such a clinical follow-up period, one may also document the response to treatment targeted at the outcome diagnosis and use this information to determine whether the target disease was present at t = 0. Response to therapy may be helpful in excluding (in the case of no response) or confirming (in the case of a beneficial effect on symptoms) the target disease. In these situations, one should be aware that response following therapy provides no definite proof of the disease, because the response could result from other factors. Similarly, lack of response does not preclude the presence of the disease at t = 0. Examples of using the response to empirical treatment to confirm a diagnosis are studies in suspected heart failure [Kelder et al., 2011].

**Partial and differential outcome verification.** Ideally, the index tests and reference standard are determined in all study participants and in a standardized manner. For various reasons, however, the reference standard may not have been performed in all patients. Such partial outcome verification might be attributable to ethical concerns or patient or physician preferences (e.g., when the reference test is considered unnecessary or too burdening, or because it is simply impossible to perform in all patients; for example, biopsy and histology as the reference standard in diagnosing cancer can only be performed in subjects with detected nodes or hot spots based on previous testing [de Groot et al., 2011a; Reitsma et al., 2009]). Partial outcome verification (i.e., partially missing outcome data) often occurs not completely at random but selectively. The reason for performing the reference standard is typically related to the test results of preceding index
tests. Such partial verification may lead to biased estimates of the accuracy of the index tests if only the selective subsample of patients in whom the reference test was executed are included in the analysis. This is known as partial verification bias, work-up bias, or referral bias [Rutjes et al., 2007]. Often researchers use a different, second best, reference test to verify the target disease presence in those subjects for whom the first, preferred reference test cannot be used [de Groot et al., 2011b]. Such differential verification will lead to bias when the results of the two reference tests are treated in the analysis as interchangeable, while both are of different quality in classifying the target disease or may even define the target disease differently. Hence, simply combining all disease outcome data in a single analysis as if both reference tests yield the same disease status does not reflect the “true” pattern of disease presence. Such an estimation of disease prevalence thus differs from what one would have obtained if all subjects had undergone the preferred reference standard. Consequently, all estimated measures of the accuracy of the diagnostic index tests will be biased; this is called differential verification bias [de Groot et al., 2011b; Reitsma et al., 2009]. Several solutions to deal with partial and differential outcome verification and its consequential bias have been proposed [de Groot et al., 2011b, 2001c]. One solution is multiple imputation of missing outcomes.

Design of Data Analysis

Objective of the Analysis

Analysis of data from multivariable diagnostic accuracy research (as opposed to test research) may serve a number of purposes: (1) to show which potential diagnostic determinants independently contribute to the estimation of the probability of disease presence (i.e., which determinants change the probability of disease presence); (2) to quantify to what extent these contributing determinants change the probability of disease presence (i.e., to estimate the relative accuracy or weights of these determinants); (3) to develop and/or validate a diagnostic model or rule to facilitate the estimation of the probability of disease given the combination of test results in individual patients in clinical practice [Moons et al., 2004a; Moons et al., 2012a].

Whether all three goals can or should be pursued depends on the motive of the study. If the aim is only to determine whether a particular test has added value or may replace another existing test, then the third goal may not be relevant. Furthermore, prior knowledge and the amount and type of study data determine whether the second and third goals should be addressed, as we will discuss next. We do not intend to provide full details
on the statistical analysis of diagnostic data. For this we refer to the statistical literature.

**Required Number of Subjects**

The multivariable character of diagnostic research creates problems for the estimation of the required number of study subjects. Power calculations do exist for test research, that is, studies aiming to estimate the diagnostic value (e.g., sensitivity, specificity, predictive values, likelihood ratios, or ROC area) of a single test or to compare the properties of two single tests [Hanley & McNeil, 1983; Simel et al., 1991]. For multivariable studies that aim to quantify the independent contribution of each test with sufficient precision, no straightforward methods to estimate the required patient number are available. Several authors have stipulated, however, that in multivariable prediction research, including diagnostic studies, for each determinant (or diagnostic test) studied at least 10 subjects are needed in the smallest category of the outcome variable to allow proper statistical modeling. In case of the typical dichotomous outcome, that is, those with or without the disease, this usually implies 10 individuals with the disease [Harrell et al., 1996; Peduzzi et al., 1996]. If the number of potential determinants is much larger than 10% of the number of diseased, the analysis tends to overestimate the accuracy of the diagnostic strategy or model. The expected number of patients with the target disease thus limits the number of determinants to be analyzed and what might be inferred from a study.

**Univariable Analysis**

Before proceeding to multivariable analyses, we recommend first performing a univariable analysis in which each individual potential determinant is related to the outcome. Biostatisticians often refer to this type of analysis as a *bivariate analysis* because the association between two variables (determinant and outcome) is studied. In diagnostic research, categorical determinants with more than two categories and continuous determinants are often dichotomized by introducing a threshold. This commonly leads to loss of information [Royston et al., 2006]. For example, dichotomizing the body temperature $> 37.5^\circ$ Celsius (C) as test-positive and $\leq 37.5^\circ$ Celsius as test-negative implies that the diagnostic implications for a person with a temperature of 38.0°C are the same as for a person with a temperature of 41°C. Second, the resulting association heavily depends on the threshold applied. This may explain why different studies of the same diagnostic test yield different associations. The aim of univariable analysis is to obtain insight into the association of each potential determinant and the presence or absence of the disease. Although it is common to only include in the multivariable
analysis the determinants that show statistical significance (P-value < 0.05), in univariable analysis this may lead to optimistic estimates of the accuracy of a diagnostic model [Harrell, 2001; Steyerberg et al., 2000; Sun et al., 1996]. This chance of “optimism” increases when the number of potential determinants clearly exceeds the “1 to 10 rule” described earlier. It is therefore recommended to use a more liberal selection criterion, for example, P < 0.20, 0.25, or an even higher threshold [Steyerberg, 2009]. The downside to this is that more determinants will qualify for multivariable analysis, requiring the need for so-called internal validation and penalization or shrinkage methods that we will discuss later in this chapter. Alternatively, univariable analyses may guide combination and clustering of determinants, ideally influenced by prior knowledge of the most important determinants. Methods have been developed to incorporate prior knowledge into the selection of predictors [Harrell, 2001; Steyerberg et al., 2004]. Finally, univariable analysis is useful to determine the number of missing values for each determinant and for the outcome, and whether these missing values are missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR).

**Multivariable Analysis**

Diagnostic practice is probabilistic, multivariable, and sequential. Consequently, a multivariable approach is the main component of the data analysis in diagnostic research. In the multivariable analysis, the probability of disease is related to combinations of multiple diagnostic determinants, in various orders. Multivariable analysis can accommodate the order in which tests are used in practice and will show which combination of tests truly contributes to the diagnostic probability estimation. To address the chronology and sequence of testing in clinical practice, the accuracy of combinations of easily obtainable determinants should be estimated first and subsequently the added value of the more burdensome and costly tests [Moons et al., 1999].

Logistic regression modeling is the generally accepted statistical method for multivariable diagnostic studies with a dichotomous outcome [Harrell, 2001; Hosmer & Lemeshow, 1989]. Other statistical methods, such as neural networks and classification and regression trees (CART), have been advocated, but these received much criticism as both often result in overly optimistic results [Harrell, 2001; Tu, 1996]. Therefore, we will focus on the use of logistic regression models for multivariable diagnostic research.

The determinants included in the first multivariable logistic regression model are usually selected on the basis of both prior knowledge and the results of univariable analysis. Also, the first model tends to concentrate on
determinants that are easy to obtain in practice. Hence, this model typically includes test results from history taking and physical examination [Moons et al., 2004a; Moons et al., 1999]. A logistic regression model estimates the log odds (logit) of the disease probability as a function of one or more predictors:

$$\log \left[ \frac{\text{probability (outcome event)}}{\text{probability (nonevent)}} \right] = \beta_0 + \beta_1 \times T_1 + \ldots \beta_n \times T_n = \text{linear predictor} \quad \text{(Eq. 1)}$$

in which $\beta_0$ is the intercept and $\beta_1$ to $\beta_n$ are regression coefficients of $T_1$ to $T_n$. $T_1$ to $T_n$ are the results of the diagnostic determinants (tests) obtained from patient history and physical examination. The sum of the intercept and the regression coefficients multiplied by the measured values of the determinants is called the linear predictor ($lp$) [Harrell et al., 1996]. A regression coefficient can be interpreted as the log odds of the outcome event relative to a nonevent per unit increase in a specific test, or in the case of a dichotomous test, the log odds of the outcome event for a positive relative to a negative test. The odds ratio can be computed as the antilog of the regression coefficient [$\exp(\beta)$]. Equation 1 can be rewritten to estimate the probability of the outcome event for an individual patient:

$$\text{Probability (disease presence)} = \frac{\exp(\beta_0 + \beta_1 \times T_1 + \ldots \beta_n \times T_n)}{1 + \exp(\beta_0 + \beta_1 \times T_1 + \ldots \beta_n \times T_n)} \quad \text{(Eq. 2)}$$

$$= \frac{1}{1 + \exp[-(lp)]}$$

The probability of absence of disease can be estimated as:

$$\text{Probability (disease absence)} = 1 - \text{probability (disease presence)} \quad \text{(Eq. 3)}$$

The next step is to remove the noncontributing determinants to obtain a reduced model with a similar diagnostic performance as the full multivariable model. Noncontributing tests are manually (one by one) excluded using the log likelihood ratio test, again at a liberal level; for example, diagnostic tests could be excluded if the significance level (P-value) exceeds, say 0.10 or 0.15. This leads to a so-called reduced model that includes only those history and physical determinants that independently contribute to the probability estimation. The regression coefficient of each determinant reflects its independent contribution (weight) to the outcome probability (see Equation 1).

The next step is to estimate the diagnostic accuracy of this reduced multivariable model. The accuracy of a model is commonly estimated by two parameters: the calibration (reliability or goodness of fit) and the discrimination [Harrell, 2001; Hosmer & Lemeshow, 1989; Steyerberg, 2009]. Calibration is
measured by the level of agreement between the disease probabilities (ranging from 0–100%) estimated by the model versus the observed disease frequencies. This is usually quantified by constructing equally large patient subgroups (say 20) after ordering of the estimated disease probabilities of all individual participants (from 0–100%) and by comparing the calculated frequency of the disease in each subgroup (in this case from those at the lowest to those at the highest 5% end of the distribution) to the number of diseased observed in each category. Good calibration means that the estimated probability of disease presence in the subgroups is similar to the observed disease frequency. The best way to examine this is by a graphical comparison. Figure 2–2 shows a calibration plot of a “reduced diagnostic history and physical model” for the diagnosis of deep vein thrombosis.

**FIGURE 2–2** Calibration plot of a reduced multivariable logistic regression model, including six determinants from patient history and physical examination to estimate the probability of the presence of DVT in 400 patients suspected of DVT. The dotted line represents the line of identity, that is, perfect model calibration. All triangles represent 10% of the patients. The triangle on the left end represents the 10% with the lowest predicted probability of disease, with the mean predicted probability (32%) on the x-axis and a somewhat lower observed prevalence of DVT (28%) in the same patients on the y-axis.
(DVT) estimated from 400 primary care patients suspected of DVT. Ideally, the slope of the calibration plot is 1 and the intercept 0. The presented model includes six patient history and physical examination determinants, taking the form of Equation 1. The calibration of this model was very good, as the predicted probabilities are very similar to the observed disease prevalence across the entire distribution. Figure 2–2 shows a slight overestimation by the model in those patients in the lower estimated disease probability range.

A common statistic used to assess whether a multivariable model shows good calibration is the goodness-of-fit test. A statistically significant ($P < 0.05$) test indicates marked differences between predicted and observed probabilities and thus poor calibration [Hosmer & Lemeshow, 1989]. This test, however, often lacks statistical power to determine important deviations from good calibration because the P-value is seldom less than 0.05 [Harrell, 2001; Hosmer & Lemeshow, 1989]. We therefore recommend that the investigator closely examines the calibration plot to determine a model’s calibration.

The discrimination of a multivariable model refers to the model’s ability to discriminate between subjects with and without the disease. This is estimated with the area under the ROC curve or the c-index (index of concordance) of the model [Hanley & McNeil, 1982; Harrell et al., 1982]. Figure 2–3 shows the ROC curve of the “reduced multivariable history and physical examination model.” A multivariable model in fact can be considered a “single” test, existing of several component tests, with the model’s estimated probability of disease presence (using Equation 2) as the “single” test result. The ROC curve exhibits the sensitivity (“true-positive rate”) and 1 – specificity (“false-positive rate”) of the model for each possible threshold in the range of “estimated probabilities.” The area under the ROC curve reflects the overall discriminative value of the model, irrespective of the chosen threshold. It exhibits the extent to which the model can discriminate between subjects with and without the target disease. The diagonal line reflects the worst model or test; for each threshold, the number of correctly diagnosed patients equals the number of false diagnoses, that is, no discriminating value and an ROC area of 0.5 (“half of the square”). In other words, the probability of a false and true diagnosis is both 50% and such a model is no better than flipping a coin. The best model is reflected by the “curve” that runs from the lower left to the upper left and upper right corners, yielding an ROC area of 1.0 (“the entire square”). Hence, the more the ROC curve is in the left upper corner—the higher the area under the curve (the closer to 1.0)—the higher the discriminative value of the model. More exactly defined, the ROC area is the probability that for each (randomly) chosen pair of one diseased and one nondiseased subject, the model estimates a higher probability for the diseased than for the
nondiseased individual [Hanley & McNeil, 1982; Harrell et al., 1982]. In our example, the ROC area of the “reduced history + physical model” was 0.70.

The next step is to extend this model by the subsequent test from the workup in our example study on DVT; this was the D-dimer assay. This allows estimation of the assay’s diagnostic value in addition to the items from history taking and physical examination. In this analysis, the same statistical procedures as just described are used. Whether the D-dimer test is a truly independent predictor is estimated again by the log likelihood ratio test [Harrell, 2001; Hosmer & Lemeshow, 1989]. Next, the calibration and discrimination of the extended model (including the “reduced history + physical model” items plus the D-dimer assay) are examined. The calibration of this extended model was good (data not shown), and the discriminatory value was high (ROC area = 0.84; Figure 2–3). Methods have been proposed to formally estimate the precision of differences between ROC areas, in this case 0.84–0.70 = 0.14, by calculating the 95% confidence interval (CI) or

FIGURE 2–3  Example of an ROC curve of the reduced multivariable logistic regression model, including the same six determinants as in Figure 2.2. The ROC area of the “reduced history + physical model” was 0.70 (95% confidence interval [CI], 0.66–0.74) and of the same model added with the D-dimer assay 0.84 (95% CI, 0.80–0.88).
P-value of this difference. In this calculation, one needs to account for the correlation between both models (“tests”) as they are based on the same subjects [Hanley & McNeil, 1983]. In our example study, the CIs did not overlap, indicating a significant added value of the D-dimer assay at the 0.05 level.

This process of model extension can be repeated for each subsequent test. Moreover, all of these analytic techniques can be used to compare the difference in the added diagnostic value of two tests separately when the aim is to choose between the two or to compare the diagnostic accuracy of various test orders. We should emphasize that the ROC area of a multivariable diagnostic model or even a single diagnostic test has no direct clinical meaning. It estimates and can compare the overall discriminative value of diagnostic models or strategies.

The DVT example exemplifies the need for multivariable diagnostic research. A comparison between models including fewer or additional tests enables the investigator to learn not only about the added value of tests but also about the relevance of moving from simple to more advanced testing in practice. It should be noted that the data analysis as outlined here only quantifies which subsequent tests have independent or incremental value in the diagnostic probability estimation and thus should be included in the final diagnostic model from an accuracy point of view. It might still be relevant to judge whether the increase in accuracy of the test outweighs its costs and patient burden. This weighing can be done formally, including a full cost-effectiveness or cost-minimization analysis accounting for the consequences and utilities of false-positive and false-negative diagnoses [Moons et al., 2012b; Vickers & Elkin, 2006]. This enters the realm of medical decision making and medical technology assessment and is not covered here.

The multivariable analysis can be used to create a clinical prediction rule that can be used in clinical practice to estimate the probability that an individual patient has the target disease given his or her documented test results. There are various examples of such multivariable diagnostic rules: a rule for diagnosing the presence or absence of DVT [Oudega et al., 2005b; Wells et al., 1997], pulmonary embolism [Wells et al., 1997], conjunctivitis [Rietveld et al., 2004], and bacterial meningitis [Oostenbrink et al., 2001]. How to derive a diagnostic rule, the ways to present it in a publication and how to enhance its use in clinical practice will be described next.

Internal Validation and Shrinkage of the Diagnostic Model

An initial prediction model commonly shows a too optimistic discrimination (ROC area relatively high, closer to 1.0) and calibration (slope close to 1.0 and intercept close to 0) when it is applied to the data from which it is derived (i.e., the derivation or development data set). The model is so-called
This means that the model's predicted probabilities will be too extreme (too high for the diseased and too low for the nondiseased) when the model is applied to new patients; calibration will be poorer and discrimination lower in daily practice [Altman et al., 2009; Moons et al., 2012b]. The amount of optimism (overfitting) in both calibration and discrimination can be estimated using so-called internal validation methods. Here internal means that no new data are used, just data from the derivation set.

The most widely used internal validation methods are the split-sample, cross-validation, and bootstrapping methods [Harrell, 2001; Steyerberg, 2009]. In the first two, part of the derivation data set (e.g., a random sample of 75% or a sample based on the time of inclusion in the study) is used for model development. The remainder (25%) is applied for estimating the model's accuracy. With bootstrapping, first a model is developed (fitted) on the full sample as described earlier. Then, multiple random samples (e.g., 100) are drawn from the full sample. On each bootstrap sample, the model is redeveloped. The calibration (slope and intercept of the calibration plot) and discrimination (ROC area) of each bootstrap model are then compared to the corresponding estimates of the bootstrap models when applied (tested) in the original full sample. These differences can be averaged, and they provide an indication of the average optimism of the bootstrap models. This average optimism in discrimination and calibration can be used to adjust the original model estimated in the full sample, that is, adjusting or shrinking the regression coefficients and ROC area. Application of the shrunken model (regression coefficients) in new patients will generally yield better (less optimistic) calibration, and the adjusted discrimination (ROC area) better approximates the discrimination that can be expected in clinical practice [Harrell, 2001; Steyerberg et al., 2001]. Bootstrapping is preferred over split-sample or cross-validation as an internal validation tool as it is more efficient; bootstrapping uses all patient data for model development and for the model validation. Importantly, all steps in the model's development, including decisions on the transformation, clustering, and re-coding of variables as well as on the selection of variables (both in the univariable and multivariable analysis) can and should be redone in every bootstrap sample [Harrell, 2001; Steyerberg et al., 2003]. Bootstrapping techniques have become widely available in standard statistical software packages, such as STATA, SAS, and S-plus. Alternative methods for shrinkage or penalizing a model for potential overfitting are the use of a heuristic shrinkage factor [Copas, 1983; van Houwelingen & LeCessie, 1990] and the use of penalized estimation methods [Harrell, 2001; van Houwelingen, 2001; Moons et al., 2004b].
Inferences from Multivariable Analysis

The lower the number of study patients and the higher the number of candidate determinants, the larger the chance of optimism of the final diagnostic model and the need for bootstrapping and shrinkage. Under certain extreme circumstances, even bootstrapping and shrinkage techniques cannot account for all optimism [Bleeker et al., 2003; Steyerberg et al., 2003]. The analysis and inferences then should be more cautious. Preferably one should then not try to achieve the third goal described previously, but rather restrict the analysis to identifying independent predictors of the presence or absence of the disease (first goal) and estimate their shrunken relative weights (second goal). If after bootstrapping and shrinkage a full model is still reported, we advise investigators to stress the need for future studies focused on confirming the observed predictor–outcome associations, and to estimate the calibration and discrimination of these predictors in new patient samples.

Prediction Rules and Scores

A diagnostic model developed to assist in setting a diagnosis in individual patients can be presented (or reported) in three ways. The most precise method is to report the original (untransformed) logistic model with the shrunken regression coefficients and corresponding discrimination and calibration of the model. This model presentation has the form of Equation 1. Readers may apply this model directly to estimate an individual patient’s probability of the disease by multiplying the patient’s test results by the corresponding coefficients, summing these up, and taking the antilog of the sum using Equation 2. This, however, requires a calculator or computerized patient record, which are not always available in clinical practice. To improve the applicability of a multivariable model in practice, one can use the (shrunken) regression coefficients to create a nomogram, as shown in Figure 2–4. This is rarely done, although the creation of a nomogram has become easy with the statistical package S-plus.

A final method to present a prediction model and to facilitate its implementation is a so-called simplified risk score or scoring rule. The original (shrunken) regression coefficients (first method, Equation 1) are then transformed to rounded numbers that are easily added together. This is commonly done by dividing each regression coefficient by the smallest regression coefficient, multiplying it by 10, and rounding this to the nearest integer. The reporting of a simplified rule must be accompanied by the observed disease frequencies across score categories, as we will show in the example that follows. This simplification of a risk score will lead to some loss of information and thus some loss in diagnostic accuracy, because the original regression coefficients are simplified and rounded. However, this
loss in precision usually does not affect clinical relevance. Ideally, the loss in precision should be minimal, with the simplified risk score as accurate as the original model but more easy to use. To allow readers to choose, we recommend that the report includes both the original untransformed model and the simplified risk score with the ROC areas.

The multivariable analysis presented earlier shows which combination of tests best predicts the presence of disease (or whether a new or alternative test improves prediction) and provides a tool to estimate an individual patient’s probability of having a specific disease. It does not quantify in what proportion of suspected patients the use of a diagnostic model or the addition of a new or alternative test will change patient management. Such a change in patient management can best be illustrated in Figure 2–1.

**FIGURE 2–4** Nomogram of a diagnostic model used to estimate the probability of DVT in suspected patients. To use this nomogram, a man (corresponding with 7 points from the “Points” scale at the top of the figure), who (obviously) does not use oral contraception (0 points), has no leg trauma (6 points), and no recent malignancy (0 points), underwent surgery in the past 3 months (4 points), has a difference in calf circumference more than 3 cm (11 points), no vein distention (0 points), and a D-dimer concentration of ≥ 500 μg/L (20 points), receives a “Total Points” score of 48. The lower two scales of the graphic show that this score corresponds to a probability of DVT of about 0.55 (or 55%).
Patients suspected of having a disease can be categorized as those with a probability of the disease low enough to exclude the diagnosis (i.e., below threshold A), those with a probability high enough to consider the disease to be present (i.e., beyond threshold B), and those in the grey area in between, where additional testing may be considered. For a new or alternative diagnostic strategy or test to have an impact on patient management, the proportion of patients that is correctly reclassified from one category to another (thus those with the disease to a higher category and those without to a lower category) should be high enough. When, for example, the addition of a new test increases the estimated probability of disease in some patients with the target disease from, for example, 80–90%, while both proportions lie above the B threshold for this particular disease, the impact in daily practice will be limited. When, however a new test correctly reclassifies many patients from the grey area to either the area below the A or above the B threshold, its impact will be much higher. Such a quantification of the reclassification of patients (through, for example, the net reclassification improvement) is increasingly being applied in prediction research [Pencina et al., 2008; Steyerberg et al., 2012]. This requires, however, definition of the thresholds A and B. This may be quite a challenge, as it typically requires reaching consensus about something rather subjective. Methods to formally quantify the optimal probability thresholds are available, but they fall beyond the scope of this text.

**External Validation**

As explained earlier, the possible optimism of a diagnostic model may be addressed by internal validation. However, external validation, using new data, is generally necessary before a model can be used in practice with confidence [Altman & Royston, 2000a; Justice et al., 1999; Reilly & Evans, 2006]. External validation is the application and testing of the model in new patients. The term *external* refers to the use of data from subjects who were not included in the study in which the prediction model was developed. So defined, external validation can be performed, for example, in patients from the same centers but from a later period than that during which the derivation study was conducted, or in patients from other centers or even another country [Justice et al., 1999; Reilly & Evans, 2006]. External validation studies are clearly warranted when one aims to apply a model in another setting (e.g., transporting a model from secondary to primary care) or in patient subgroups that were not included in the development study (e.g., transporting a model from adults to children) [Knottnerus, 2002a; Oudega et al., 2005a].

Too often, researchers use their data only to develop their own diagnostic model, without even mentioning—let alone validating—previous models.
This is unfortunate as prior knowledge is not optimally used. Moreover, recent insights show that in the case where a prediction (diagnostic or prognostic) model performs less accurately in a validation population, the model can easily be adjusted based on the new data to improve its accuracy in that population [Moons et al., 2012b; Steyerberg et al., 2004]. For example, the original Framingham coronary risk prediction model and the Gail breast cancer model were adjusted based on later findings and validation studies [Costantino et al., 1999; Grundy et al., 1998]. An adjusted model will then be based on both the development and the validation data set, which will further improve its stability and applicability to other populations. The adjustments may vary from parsimonious techniques such as updating the intercept of the model for differences in outcome frequency, via adjusting the originally estimated regression coefficients of the determinants in the model, to even adding new determinants to the model. It has been shown, however, that simple updating methods are often sufficient and thus preferable to the more extensive model adjustments [Janssen et al., 2008 & 2009; Steyerberg et al., 2004].

With these advances, the future may be one in which prediction models—provided that they are correctly developed—are continuously validated and updated if needed. This resembles cumulative meta-analyses in therapeutic research. Obviously, the more diverse the settings in which a model is validated and updated, the more likely it will generalize to new settings. The question arises about how many validations and adjustments are needed before it is justifiable to implement a prediction model in daily practice. Currently there is no simple answer. “Stopping rules” for validating and updating prediction models should be developed for this purpose.

APPLICATION OF STUDY RESULTS IN PRACTICE

Why are prediction models constantly used in, for example, weather forecasting and economics (albeit with varying success), while they still have limited application in medicine? There are several potential explanations. First, prediction models are often too complex for daily use in clinical settings that are not supported by computer technology. This may improve with the introduction of computerized patient records but also may require a change in attitude by practicing physicians. Second, because diagnostic (and prognostic) models often are not routinely validated in other populations, clinicians may not—and perhaps should not—trust the probabilities provided by these models. External validation studies as described earlier in the chapter are still scarce. Even less frequently are models validated or tested for their ability to change clinicians’ decisions, not to mention their
ability to improve a patient’s prognosis [Reilly & Evans, 2006; Stiell et al., 1995]. There are no formal criteria to judge the generalizability of diagnostic study results, but a few rules of thumb can be given. Generalizability of a diagnostic model is first and foremost determined by its use in the appropriate domain of patients suspected of having the target disease. Second, it is commonly determined by the setting (primary, secondary, tertiary care) in which the model was developed and perhaps validated. For example, particular symptoms or signs presented by patients in an academic hospital may be less relevant in patient populations from a general hospital or from primary care and vice versa [Knottnerus, 2002a]. This has been shown, for example, for extrapolation of a diagnostic rule for DVT developed in secondary care patients to primary care patients [Oudega et al., 2005a]. Third, generalizability is determined by the tests included in the final model. For example, the inclusion of particular advanced tests, such as spiral CT scanning, may lead to a limited applicability of the model to other patient populations or settings.

A final reason why diagnostic models are often not applied in daily practice is that clinicians may find it difficult to include explicit predicted probabilities in their decision making; many doctors are reluctant to accept that a simplified mathematical formula replace their clinical experience, skills, and complicated diagnostic reasoning in everyday patient care. The latter opinion clearly is a misunderstanding. Diagnostic rules are tools that should be used to aid physicians in their daily tasks, indeed, to help them cope with their complicated diagnostic challenges. Such tools are not meant to be a substitute for clinical experience and skills, but to strengthen them.

WORKED-OUT EXAMPLE

Recognition and ruling out of DVT is difficult based on history taking and physical examination alone. An adequate diagnosis in patients presenting with symptoms suggestive of DVT (usually a painful, swollen leg) is crucial because of the risk of potentially fatal pulmonary embolism when DVT is not adequately treated with anticoagulants. False-positive diagnoses also should be avoided because of the bleeding risk associated with anticoagulant therapy. The serum D-dimer test clearly improves the accuracy of diagnosing and ruling out DVT in suspected patients. Algorithms, including clinical assessment (i.e., signs and symptoms) and D-dimer testing are available that are widely applied in clinical practice and recommended in current guidelines. The most famous of these, the Wells rule, was developed and validated in secondary care settings [Wells et al., 1997]. Research
demonstrated that the Wells rule cannot adequately rule out DVT in patients suspected of DVT in primary care as too many (16%) patients in the low-risk category (Wells score below 1) still had DVT [Oudega et al., 2005a].

The goal of the study presented here (see Box 2–8), was to develop the optimal diagnostic strategy, preferably by way of a diagnostic rule, to be applied in the primary care setting [Oudega et al., 2005b].

Theoretical Design

The research question was: “Which combination of diagnostic determinants best estimates the probability of DVT in patients suspected of having DVT in primary care?”

Determinants considered included findings from history taking and physical examination as well as the D-dimer test result. The occurrence relation can be summarized as:

\[ P(DVT) = f(T_1, T_2, T_3, \ldots, T_n) \]
where $T_1, \ldots, T_n$ refer to all potential diagnostic determinants studied (in total 17).

The domain of the study consisted of patients presenting to primary care with symptoms suggestive of DVT.

**Design of Data Collection**

Data were collected cross-sectionally. Participating primary care physicians were asked to include all patients in whom the presence of DVT was suspected during an inclusion period of 17 months. All 17 diagnostic determinants and the reference standard were assessed in all included patients. Thus, the time dimension of data collection was zero, a census (and no sampling) approach was taken, and the study was observational (and not experimental).

The inclusion criterion was phrased as “all patients aged 18 years or older in whom the primary care physician suspected deep vein thrombosis,” while in the information forwarded to the primary care physician, suspicion of DVT was explicitly defined as at least one of the following symptoms or signs of the lower extremities: swelling, redness, and/or pain. Exclusion criteria included a duration of the symptoms exceeding 30 days and suspicion of pulmonary embolism. In total, 110 primary care physicians in three regions in the central part of the Netherlands, each served by one hospital, were involved.

All items from history and physical examination were recorded in the case record form by the patient’s primary care physician. The D-dimer test and the reference standard (real time B-mode compression ultrasonography) were performed in the adherent hospital. In patients with a normal compression ultrasonography, the procedure was repeated after 7 days to definitely rule out DVT. The diagnostic determinants under study and the result from the reference standard were recorded in all 1,295 included patients.

**Design of Data Analysis**

After univariable analysis, a multivariable logistic regression analysis was done including all 16 findings from history taking and physical examination in the model to determine which of these independently contributed to the presence or absence of DVT. Model reduction was performed by excluding variables from the model with a P-value > 0.10 based on the log likelihood ratio test. Subsequently, the D-dimer test was added to the reduced “history + physical” model to quantify its incremental value,
which resulted in the final model. The calibration and ROC area of both models (with and without D-dimer) were estimated. Bootstrapping techniques, repeating the entire modeling process, were used to internally validate the final model and to adjust the estimated performance of the model for optimism. The model’s performance obtained after bootstrapping was considered to approximate the expected performance in similar future patients. To construct an easily applicable diagnostic rule, the regression coefficients of the variables in the final model were transformed to integers according to their relative contributions (quantified through the regression coefficients) to the probability estimation. Finally, after estimating the score for each patient, the absolute percentages of correctly diagnosed patients across score categories were estimated. One hundred and twenty-seven subjects had missing values for one or more tests under study. Per predictor, on average, 2–3% of the values were missing. As data were not MCAR, deleting subjects with a missing value would lead not only to a loss of statistical power but also to biased results. To decrease bias and increase statistical efficiency, the missing values were imputed.

Results and Implications

Of the 1,295 patients included, 289 had DVT (prevalence 22%). An abnormal D-dimer level was by far the strongest determinant of the presence of DVT (univariable odds ratio of 35.7; 95% CI, 13.3–100.0). In multivariable analysis, 7 of the history and physical examination items were independent predictors of DVT: male gender, use of oral contraceptives, presence of malignancy, recent surgery, absence of leg trauma, vein distension, and a difference in calf circumference between the two legs of 3 cm or more. The ROC of this model was 0.68 (95% CI, 0.65–0.71). The multivariable model including these 7 determinants plus the D-dimer test had an ROC area of 0.80 before and 0.78 (95% CI, 0.75–0.81) after bootstrapping and shrinkage. This indicates a substantial added value. The odds ratio of the D-dimer assay (after shrinkage) was 20.3 (8.3–49.9). The calibration plot—after bootstrapping—of the final model showed good calibration; the P-value of the goodness of fit test was 0.56.

The final, untransformed model after shrinkage was:

\[
\text{Probability of DVT} = \frac{1}{1 + \exp(-(-5.47 + 0.59 \times \text{male gender} + 0.75 \times \text{OC use} + 0.42 \times \text{presence of malignancy} + 0.38 \times \text{recent surgery} + 0.60 \times \text{absence of leg trauma} + 0.48 \times \text{vein distension} + 1.13 \times \text{calf difference} \geq 3 \text{cm} + 3.01 \times \text{abnormal D-dimer})}
\]
To facilitate application of this model in daily practice, the following simplified scoring rule was derived:

\[
\text{Score} = 1 \times \text{male gender} + 1 \times \text{oral contraceptive use} + 1 \times \text{presence of malignancy} + 1 \times \text{recent surgery} + 1 \times \text{absence of leg trauma} + 1 \times \text{vein distension} + 2 \times \text{difference in calf circumference} \geq 3 \, \text{cm} + 6 \times \text{abnormal D-dimer test}
\]

The score ranged from 0–13 points, and the ROC area of the simplified rule was also 0.78. Table 2–1 shows the number of participants and probability of DVT in different categories of the risk score.

As an example, a woman using oral contraceptives who was without a leg trauma but had vein distension and a negative D-dimer test would receive a score of 3 (0 + 1 + 0 + 0 + 1 + 1 + 0 + 0), corresponding with a very low estimated probability of DVT of 0.7%.

It was concluded from the study that a simple diagnostic algorithm based on history taking, physical examination, and D-dimer testing can be helpful in safely ruling out DVT in primary care and thus would reduce the number of unnecessary referrals for suspected DVT.

Later, the accuracy of this simplified rule was externally validated in three regions in the Netherlands [Büller et al., 2009]. This study showed that among DVT-suspected patients not referred for ultrasonography in daily practice because of a risk score of \( \leq 3 \), the proportion with a diagnosis of DVT or pulmonary embolism within 3 months was indeed low (1.4%). The rule has been included in the current primary care clinical guideline on suspected DVT in the Netherlands.

<table>
<thead>
<tr>
<th>Risk</th>
<th>Score Range</th>
<th>Number of Patients (%)</th>
<th>Number of Patients with DVT Present (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very low</td>
<td>0–3</td>
<td>293 (23%)</td>
<td>2 (0.7%)</td>
</tr>
<tr>
<td>Low</td>
<td>4–6</td>
<td>66 (5%)</td>
<td>3 (4.5%)</td>
</tr>
<tr>
<td>Moderate</td>
<td>7–9</td>
<td>663 (51%)</td>
<td>144 (21.7%)</td>
</tr>
<tr>
<td>High</td>
<td>10–13</td>
<td>273 (21%)</td>
<td>140 (51.3%)</td>
</tr>
<tr>
<td>Total</td>
<td>0–13</td>
<td>1,295</td>
<td>289 (22.0%)</td>
</tr>
</tbody>
</table>